Vom Fachbereich für Mathematik und Informatik
der Technischen Universität Braunschweig
genehmigte Dissertation
zur Erlangung des Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
von

Dipl.-Inform. Axel Böger

Skalierbare Gruppenkommunikationsunterstützung für ATM-Netze

1. Referentin:	Prof. Dr. M. Zitterbart, Universität Karlsruhe
2. Referent:	Prof. Dr. O. Spaniol, RWTH Aachen
Eingereicht am:	22.08.2001
Mündliche Prüfung:	19.10.2001

Kurzfassung

In dieser Arbeit wird ein Ansatz vorgestellt, der eine **ska**lierbare **G**ruppenkommunikationsunterstützung für **A**TM-**N**etze (SkaGAN) ermöglicht. Die herkömmliche rechnergestützte Kommunikation findet zwischen einem Sender und einem Empfänger statt. Die Gruppenkommunikation erweitert diese Form und erlaubt einer Gruppe von Rechnern untereinander zu kommunizieren. Diese Arbeit legt dabei den Fokus auf die ATM-Technologie (Asynchronous Transfer Mode), die keine akzeptable Gruppenkommunikationsunterstützung anbietet.

ATM-Netze können sehr hohe Datenraten unterstützen und weitere Anforderungen erfüllen, (z. B. eine niedrige Verzögerung) und eignen sich damit besonders für multimediale Anwendungen. Da diese Anwendungen häufig auf der Gruppenkommunikation basieren, sollte auch für ATM-Netze eine effektive Unterstützung hierfür vorhanden sein.

Heutzutage wird ATM hauptsächlich in Backbone-Netzen eingesetzt, womit sich auch diese Arbeit auseinandersetzt. Der Schwerpunkt bei SkaGAN ist die Skalierbarkeit in Bezug auf Netzwerk- und Gruppengröße. Für den Bereich der lokalen ATM-Netze wird ebenfalls ein Lösungsvorschlag präsentiert, der eine Lastverteilung aktiver Gruppenteilnehmer auf mehrere Server beinhaltet.

Der Lösungsansatz von SkaGAN für ATM-Weitverkehrsnetze orientiert sich an dem PNNI-Routingprotokoll und basiert auf einem hierarchischen Schema. Für die Verwaltung der Gruppen wird eine Baumhierarchie eingesetzt, die eine erhebliche Reduktion des Signalisierungsaufwandes und eine gute Skalierbarkeit ermöglicht. Für den Datentransfer zwischen den Gruppenteilnehmern wird ebenfalls eine Baumstruktur eingesetzt, die sich dynamisch an Änderungen in den Gruppen anpassen kann. Dabei wird die Anzahl der benötigten Zwischensysteme möglichst gering gehalten und die Lokalität der Teilnehmer berücksichtigt. Damit konnte auch in diesen Bereich eine gute Skalierbarkeit bei der Gruppenkommunikation erreicht werden.

Danksagung

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Betriebssysteme und Rechnerverbund der Universität Braunschweig. Besonderer Dank gilt Frau Prof. Dr. Martina Zitterbart, die diese Arbeit betreut und die Durchführung in ihrer Forschungsgruppe ermöglicht hat.

Ich danke Herrn Prof. Dr. Otto Spaniol von der RWTH Aachen für die freundliche Übernahme des Korreferats.

Bedanken möchte ich mich ebenfalls bei meinen Kollegen des Instituts für Betriebssysteme und Rechnerverbund für die angenehme Zusammenarbeit. Besonders sei Herrn Jörg Diederich gedankt für Anmerkungen und Korrekturen an dieser Ausarbeitung und Herrn Kai Krasnodembski für Diskussionen über ATM und Ratschläge zum Simulationswerkzeug OpNet.

An dieser Stelle sei außerdem allen Studenten gedankt, die durch ihre Studienarbeiten, Diplomarbeiten und Hiwi-Tätigkeiten zur Umsetzung dieser Arbeit beigetragen haben. Zu erwähnen sind in diesem Zusammenhang (in alphabetischer Reihenfolge) Marc Anwander, Imed Bouazizi, Thilo Fickel, Alexander Franke, Peter Grimsehl, Christian Meier, Daniel Müller und Karsten Schubert.

Inhaltsverzeichnis

1. Einleitung

Die herkömmliche rechnergestützte Kommunikation im Internet und auch in anderen Netzwerken, wie z. B. Telefonnetzen, findet zwischen einem Sender und einem Empfänger statt. In den letzten Jahren sind auf diesem Gebiet eine Reihe von Fortschritten erzielt worden, die sich damit befassen, eine Gruppe von Rechnern untereinander kommunizieren zu lassen. In der Analogie zum Telefon entspräche das sogenannten Konferenzschaltungen, die das Telefonieren mehrerer Teilnehmer ermöglichen. Diese Form der rechnergestützten Kommunikation mit mehr als einem Sender und einem Empfänger wird als Gruppenkommunikation bezeichnet [1].

Die Gruppenkommunikation entspricht der zwischenmenschlichen Kommunikation von mehr als zwei Personen, z. B. bei einer Diskussion. Es ist in vielen Fällen auch wesentlich effizienter einer gesamten Gruppe etwas mitzuteilen, als dies für jeden Teilnehmer einzeln zu tun. Daher ist das gleichzeitige Erreichen aller Gruppenteilnehmer ein bedeutender Vorteil gegenüber der herkömmlichen Punkt-zu-Punkt-Kommunikation.

Aus der Entsprechung der rechnergestützten Gruppenkommunikation in der zwischenmenschlichen Kommunikation leiten sich viele Anwendungsgebiete für die Gruppenkommunikation ab. Hier ist vor allem die kooperative Arbeit zu nennen, wie z. B. beim verteilten Arbeiten in einem Team. Teilnehmer an verschiedenen Standorten arbeiten gemeinsam und rechnergestützt an einem Projekt. Dies wird in der Literatur allgemein als CSCW (Computer Supported Cooperative Work) oder auch als Groupware bezeichnet. Beispiele für solche Anwendungen sind Videokonferenzen mit einem gemeinsamen Whiteboard und weiteren Hilfsprogrammen. Ein ähnlicher Anwendungsbereich ist das Open Distance Learning, worunter das ortsunabhängige Lernen im Bereich der Aus- und Weiterbildung verstanden wird. Teilnehmer an verschiedenen entfernten Orten nehmen dabei zeitgleich an Lehrveranstaltungen oder Fortbildungen aus einem anderen Ort teil. Ein weiteres Anwendungsbeispiel in diesem Bereich sind interaktive Spiele.

Andere Anwendungen der Gruppenkommunikation sind sogenannte Push-Technologien. Das ist eine Form der rechnergestützten Informationsverteilung, z. B. die Übertragung von Wetterdaten oder Börsenkursen. Im Bereich der verteilten Simulationen und des verteilten Rechnens wird die rechnergestützte Gruppenkommunikation ebenfalls eingesetzt. Eine Übersicht der Möglichkeiten heutiger Gruppenkommunikationsanwendungen gibt Tabelle 1.1 [2].

Für die Unterstützung dieser hier vorgestellten Anwendungen ist eine entsprechende Unterstützung im Rechnernetz notwendig. Die größte Akzeptanz hat in diesem Bereich IP Multicast [3, 4, 5] gefunden, das eine Gruppenkommunikationsunterstützung im Internet ermöglicht. Für nicht IP-basierte Netze gibt es hingegen kaum oder nur ei-

	Realzeit	Nicht-Realzeit
Multimedia	CSCW Videokonferenz Video-Server Internet-Audio Multimedia Ereignisse	Replikation: Video- und Web-Server Internet-Kiosk Inhalteverteilung: Intranet und Internet
Daten	Börsenkurse News Verteilung interaktive Spiele Whiteboard	Datenbeförderung: Server-Server Server-Endsystem Datenbank-Replikation Software Verteilung

Tabelle 1.1.: Gruppenkommunikationsanwendungen.

ne unzureichende Gruppenkommunikationsunterstützung. Eine dieser nicht IP-basierten Netzwerktechnologien ist ATM (Asynchronous Transfer Mode) [6, 7, 8], das keine akzeptable Gruppenkommunikationsunterstützung anbietet.

ATM-Netze sind in der Lage sehr hohe Datenraten zu unterstützen und weitere Anforderungen, wie z. B. eine niedrige Verzögerung, erfüllen zu können. Damit eignen sich ATM-Netze besonders für multimediale Anwendungen. Da diese Anwendungen häufig auf der Gruppenkommunikation basieren, sollte auch für ATM-Netze eine effektive Unterstützung hierfür vorhanden sein.

1.1. Problemstellung und Anforderungen

Die ATM-Technologie ist von Telefonunternehmen konzipiert worden, und unterscheidet sich grundlegend von den im Internet verwendeten Technologien. ATM-Netze sind verbindungsorientiert, während IP nur einen verbindungslosen Vermittlungsdienst anbietet. Darüber hinaus reserviert ATM benötigte Ressourcen im Netz vor Kommunikationsbeginn. Im Gegensatz dazu werden im Internet keine harten Garantien hinsichtlich der benötigten Ressourcen gegeben.

ATM wird heutzutage hauptsächlich in Backbone-Netzen, bzw. im Weitverkehrsbereich eingesetzt. Die Etablierung von ATM bis hin zum Endsystem hat keine Akzeptanz gefunden, da andere Technologien, wie z. B. Fast Ethernet, im lokalen Bereich kostengünstigere und für die meisten Anwendungen ebenso effiziente Alternativen darstellen.

Die meisten Forschungsergebnisse im Bereich Gruppenkommunikation (insbesondere IP Multicast) und ATM behandeln den Bereich der lokalen ATM-Netze. Lösungsansätze für ein ATM-Weitverkehrsnetz sind dagegen kaum vorhanden. Im Gegensatz zu einem lokalen Netz stellt ein Weitverkehrsnetz andere Anforderungen:

- Skalierbarkeit in Bezug auf:
  - Netzwerkgröße und -topologie
  - Gruppengröße (Anzahl der Teilnehmer in einer Gruppe)
  - Anzahl der Gruppen
  - Gruppendynamik, Änderung der Teilnehmer einer Gruppe
- Ausfallsicherheit und Zuverlässigkeit der Komponenten
- Interoperabilität mit anderen Gruppenkommunikationsprotokollen

Diese Anforderungen werden zwar auch an lokale Netze gestellt, können dort aber mit einfacheren Mitteln gelöst werden. So ist z. B. die Gruppengröße in lokalen Netzen durch die Rechneranzahl begrenzt und die Netzwerktopologie ist in der Regel sternförmig. Damit erleichtern lokale ATM-Netze eine Lösung der Problematik, da hier von einem klar begrenzten Einsatzgebiet ausgegangen werden kann.

In einem ATM-Weitverkehrsnetz kann zwar im gesamten Netz dieselbe Basistechnologie vorausgesetzt werden, allerdings können keine Annahmen über die Größe des Netzes und der Menge der angeschlossenen Systeme gemacht werden. Daher stellt die Anforderung ‚Skalierbarkeit', die für die Gruppenkommunikation sowieso ein sehr wichtiger Aspekt ist, für die hier vorliegende Arbeit die eigentliche Zielproblematik dar. Die Anforderungen an die Ausfallsicherheit der Komponenten werden hier nur zum Teil berücksichtigt. Es werden Konzepte vorgeschlagen, wie die Ausfallsicherheit des hier vorgestellten Ansatzes erhöht werden kann, aber keine weitergehenden Ausführungen hierzu unternommen. Im Bereich der Interoperabilität mit anderen Gruppenkommunikationsprotokollen, bzw. Multicast-Routingprotokollen existieren eine Reihe von Lösungsvorschlägen [9, 10, 11]. Daher wird hierzu auf die bestehenden Arbeiten verwiesen.

Mit dem in dieser Arbeit vorgestellten Lösungsansatz für eine **ska**lierbare **G**ruppenkommunikation über **A**TM-**N**etzen (SkaGAN) sollen Gruppenkommunikationsanwendungen, die in der Regel auf IP Multicast basieren, effizient in ATM-Netzen unterstützt werden können. Hierzu ist es erforderlich, den Anwendungen einen zu IP Multicast kompatiblen Transportdienst anzubieten. Dafür müssen die wichtigsten Merkmale von IP Multicast emuliert werden. Das ist zum einen die verbindungslose Kommunikation und zum anderen das Konzept der Gruppenadressen. Eine Gruppenadresse identifiziert die Menge aller Empfänger einer Gruppe. Ein Sender kann an die Gruppenadresse senden, ohne weitere Kenntnisse über die Empfänger zu haben.

Diese Merkmale von IP Multicast stehen im Kontrast zu ATM. ATM ist verbindungsorientiert, d. h. die Daten werden erst übertragen, nachdem eine Verbindung zwischen den Teilnehmern aufgebaut worden ist. Des Weiteren initiiert die Quelle einer Verbindung deren Aufbau. Das hat die Konsequenz, dass die Quelle alle angeschlossenen Teilnehmer kennen muss. Das Konzept der Gruppenadressen ist somit nicht direkt auf ATM umsetzbar.

Auf eine Integration von SkaGAN in die ATM-Technologie und den bestehenden ATM-Standards ist verzichtet worden. Änderungen in den ATM-Standards würden ebenfalls Änderungen bei den ATM-Komponenten erfordern. Da diese Voraussetzung nicht

als realistisch angesehen werden kann, setzt SkaGAN oberhalb von ATM an und baut auf der ATM-Technologie auf, erfordert aber dort keine weiteren Änderungen. Dennoch ist es möglich, den hier vorgestellten Ansatz in die ATM-Technologie zu integrieren, was zu einer Reihe von Vereinfachungen und Synergieeffekten führen würde.

1.2. Gliederung und Ergebnisse der Arbeit

In Kapitel 2 werden zunächst die Grundlagen erläutert, auf denen diese Arbeit beruht. Hierzu gehört ein Überblick über die rechnergestützte Gruppenkommunikation und insbesondere über IP Multicast. Anschließend werden die technischen Grundlagen von ATM behandelt und wie eine Emulation von IP über ATM ermöglicht wird. Das Kapitel endet mit dem von der IETF vorgeschlagenen Standard zur Unterstützung von IP Multicast über ATM, der sich aber vornehmlich auf den Bereich der lokalen Netze beschränkt.

Einen Überblick über aktuelle Forschungsarbeiten auf dem Gebiet Gruppenkommunikation über ATM gibt Kapitel 3. Es werden zunächst die Kriterien vorgestellt, anhand derer die untersuchten Arbeiten beurteilt werden. Anschließend werden die einzelnen Ansätze kurz vorgestellt und nach den Kriterien untersucht. Bei allen untersuchten Arbeiten zeigen sich Defizite in der Skalierbarkeit. Im Bereich des Datentransports und der Verwaltung von Gruppen können die meisten Ansätze nur eine eingeschränkte Skalierbarkeit vorweisen. Am Ende des Kapitels befindet sich ein Vergleich der präsentierten Ansätze. Hier ist auch der eigene Ansatz (SkaGAN) aufgenommen und detailliert dargelegt, welche Anforderungen der eigene Ansatz erfüllen kann und soll.

Das anschließende Kapitel 4 beschreibt die Vorgehensweise, mit der das Problem der Gruppenkommunikation in ATM-Netzen angegangen worden ist. Dabei ist das Gesamtproblem in zwei Problemfelder (lokalen und globalen) untergliedert, die jeweils weitestgehend getrennt voneinander gelöst worden sind. Beide Teile bauen auf den gleichen verwendeten Netzwerkkomponenten auf und benutzen zur Kommunikation ein einheitliches Nachrichtenpaketformat in Verbindung mit einem Transportprotokoll, die beide in Kapitel 4 beschrieben werden.

Ein Teil von SkaGAN behandelt den Bereich der lokalen Netze und wird in Kapitel 5 beschrieben. Insbesondere wird hier auf ein Verfahren zur Lastverteilung eingegangen, das eine erhöhte Skalierbarkeit in lokalen ATM-Netzen ermöglicht. Damit können gegenüber bisherigen Ansätzen eine erhöhte Anzahl lokaler Teilnehmer aktiv an Gruppen partizipieren, ohne dass Einschränkungen in der Übertragungsqualität von den Teilnehmern hingenommen werden müssen.

Den Kern von SkaGAN beinhaltet Kapitel 6. Hier wird ein Schema für eine Gruppenkommunikationsunterstützung in ATM-Weitverkehrsnetzen vorgestellt. Der erste Teil beschäftigt sich mit der Gruppenverwaltung. Es wird ein hierarchisches Schema vorgestellt, das eine gute Skalierbarkeit und Ausfallsicherheit bietet. Die Gruppenverwaltung ist mit dem hierarchischen Ansatz auch gut an das Routing in ATM-Netzen angepasst. Basierend auf der Verwaltung wird ein Konzept für den Datentransfer entwickelt. Das Konzept beruht auf einer Baumstruktur, die dynamisch an Gruppenänderungen anpassbar ist. Dieser Basisansatz wird noch erweitert, um eine bessere globale Lastverteilung

zwischen den Komponenten zu ermöglichen. Dadurch werden vor allem Gruppen mit vielen Teilnehmern besser unterstützt. Das Kapitel endet mit einer Leistungsbewertung, bei der die Vor- und Nachteile des entwickelten Konzeptes gegenübergestellt und ein Nachweis über die Funktionsfähigkeit des Ansatzes erbracht werden. Es kann gesagt werden, dass der Ansatz die an ihn gestellten Anforderungen erfüllt und eine gute Skalierbarkeit aufweist.

Eine Zusammenfassung und einen Ausblick enthält Kapitel 7, das den vorgestellten Ansatz von SkaGAN noch einmal rekapituliert. Im Anhang werden das verwendete Simulationswerkzeug OpNet, das realisierte Simulationsmodell, und die verwendeten Nachrichtenformate für den Informationsaustausch beschrieben.

2. Grundlagen

Dieses Kapitel erläutert die Grundlagen, die für das Verständnis der folgenden Kapitel vorausgesetzt werden. Die Inhalte sind in vier Teile gegliedert, die die Grundlagen in den Bereichen Gruppenkommunikation, ATM, IP über ATM und IP-Multicast über ATM beschreiben.

Das erste Unterkapitel 2.1 charakterisiert den Begriff Gruppenkommunikation. Der zweite wichtige Bereich ist die ATM-Technologie, die in Unterkapitel 2.2 beschrieben wird. Die Nutzung von ATM als Übertragungstechnik für das Internet behandelt Unterkapitel 2.3 und das letzte Unterkapitel 2.4 beschreibt den MARS-Ansatz (Multicast Address Resolution Server), der von der IETF als Standard für die IP-Multicast Emulation über ATM vorgesehen ist.

2.1. Gruppenkommunikation

Als Gruppenkommunikation [1, 2] wird der Datenaustausch zwischen einer Gruppe von Rechnern bezeichnet. Eine Gruppe besteht dabei aus zwei oder mehr Teilnehmern, die geografisch beliebig verteilt sein können. Gruppenkommunikation ist ein allgemeiner Oberbegriff, der verschiedene Ausprägungen haben kann:

Unicast: Das entspricht der traditionellen Punkt-zu-Punkt-Kommunikation (1:1-Kommunikation) bei der genau eine Datenquelle und ein Empfänger existieren. Mit dieser Kommunikationsform stellt Unicast in Bezug auf die Gruppenkommunikation eine besondere Ausprägung dar. Bei einer Unicast-Kommunikation wird häufig auch von einem bidirektionalen Datenverkehr ausgegangen.

Multicast: Bei der Multicast-Kommunikation handelt es sich um eine Erweiterung der Unicast-Kommunikation. Eine Datenquelle sendet seine Daten an mehrere Empfänger (1:n-Kommunikation). Die Kommunikation ist unidirektional, nur die Datenquelle kann senden.

Concast: Hierbei handelt es sich um das entgegengesetzten Vorgehen im Vergleich zu Multicast. Eine Kommunikation wird als Concast bezeichnet, wenn mehrere Sender an einen einzigen Empfänger Daten senden (m:1-Kommunikation). Auch hier ist der Datenverkehr unidirektional.

Multipeer: Das ist die allgemeinste Form der Gruppenkommunikation, bei der mehrere

Sender an mehrere Empfänger Daten versenden (m:n-Kommunikation). Die Teilnehmer innerhalb einer Gruppe können sowohl Daten senden als auch empfangen.

Es existieren noch weitere Kommunikationsformen, unter anderem Anycast und Broadcast. Anycast ist eine Form des Unicast und macht sich ebenfalls das Gruppenkonzept zu nutzen. Die Daten werden dabei aber nur von einem Mitglied der Gruppe empfangen. Welches Mitglied ausgewählt wird, hängt dabei vom Anycast-Mechanismus ab. Als Broadcast wird das Senden von Daten an alle Teilnehmer im Netz verstanden. Dieses Konzept findet vor allem in Shared-Media-Netzen wie Ethernet oder drahtlosen Netzen Anwendung (z. B. bei der Umwandlung von IP-Adressen in MAC-Adressen, wie bei ARP [12]), ist aber in großen drahtgebundenen Netzen von untergeordneter Bedeutung.

2.1.1. Eigenschaften von Gruppen

Eine Gruppe von Teilnehmern kann nach folgenden Eigenschaften differenziert werden:

Offenheit: Es wird zwischen offenen und geschlossenen Gruppen unterschieden. An offene Gruppen können beliebige Quellen Daten senden. Die Datenquelle muss dabei nicht Mitglied der Gruppe sein. In geschlossenen Gruppen ist das eingeschränkt, nur die Datenquellen, die auch Mitglied in der Gruppe sind, können an die Gruppe senden.

Dynamik: Bei der Dynamik werden statische und dynamische Gruppen unterschieden. In statischen Gruppen ist die Zusammensetzung der Mitglieder einer Gruppe vorgegeben und ändert sich während ihres Bestehens nicht. Bei dynamischen Gruppen kann sich die Zusammensetzung im Laufe der Zeit ändern.

Lebensdauer: Bei der Lebensdauer wird zwischen permanenten und transienten Gruppen unterschieden. Transiente Gruppen existieren nur solange, wie auch Mitglieder in der Gruppe sind. Das bedeutet auch, dass die Gruppe erst durch den Beitritt des ersten Gruppenteilnehmers existiert. Im Gegensatz hierzu existieren permanente Gruppen unabhängig von der Anzahl der Mitglieder.

Die hier vorgestellten Eigenschaften können miteinander kombiniert werden, bei IP-Multicast sind z. B. die Gruppen offen und transient.

2.1.2. Probleme bei der Gruppenkommunikation

Die Kommunikationsform der Gruppe ermöglicht viele neue Anwendungsformen, es gibt aber auch eine Reihe von Problemen, die bei der Gruppenkommunikation beachtet und gelöst werden müssen:

Sicherheit: Hierbei handelt es sich um die Verschlüsselung von Daten und Authentifizierung von Gruppenmitgliedern. Dadurch, dass eine Menge von Teilnehmern in die Gruppenkommunikation involviert ist, können Mechanismen für die Punkt-zu-Punkt-Kommunikation in Bezug auf die Sicherheit nur bedingt angewendet werden (z. B. Schlüsselverteilung).

Zuverlässigkeit: Zuverlässigkeit bedeutet bei der Unicast-Kommunikation die korrekte Auslieferung aller gesendeten Daten beim Empfänger. Diese Zuverlässigkeit wird durch Kontrolldaten überwacht. Bei großen Gruppen können die Kontrolldaten einen erheblichen zusätzlichen Verkehr verursachen und entsprechend viele Ressourcen verbrauchen. Daher benötigt die Gruppenkommunikation andere Konzepte für die Zuverlässigkeit.

Bei der Zuverlässigkeit wird hierzu zwischen unzuverlässig, halbzuverlässig und zuverlässig unterschieden. Der halbzuverlässige Gruppendienst stellt eine Zwischenstufe dar. Hierbei wird versucht, mit einer gewissen Wahrscheinlichkeit zu garantieren, dass mindestens eine Teilmenge die Daten zuverlässig erhält, aber nicht alle Teilnehmer.

Fluss- und Staukontrolle: Die Regulierung des Datenflusses, z. B. durch fensterbasierte oder ratenbasierte Verfahren, erweist sich bei der Gruppenkommunikation als schwierig, besonders wenn große, räumlich verteilte oder heterogene Gruppen unterstützt werden sollen. Dies gilt ebenso für die Staukontrolle, insbesondere wenn sich die Netzauslastung in verschiedenen Netzbereichen stark unterscheidet.

Verwaltung: Die Verwaltung einer Gruppe stellt eine weiteres wichtiges Problem dar. Da nicht mehr ein einzelner Teilnehmer angesprochen wird, ist ein anderes Adressierungsschema notwendig.

Grundsätzlich gibt es hierfür die Möglichkeiten der Adressierung über Teilnehmerlisten oder der Adressierung über spezielle Gruppenadressen. Bei Teilnehmerlisten wird die Gruppe aus einer Menge von Unicast-Adressen zusammengefasst. Die Gruppenadresse stellt ein anderes Konzept dar, bei dem sich die Teilnehmer für die Gruppenadresse, die die Gruppe identifiziert, registrieren. Dadurch können die Empfänger nicht mehr vom Sender unterschieden werden, es entfällt aber auch das mehrfache Aussenden der Daten beim Sender.

Bei Gruppenadressen kann noch zwischen zentraler und dezentraler Verwaltung unterschieden werden, je nachdem ob eine authorisierte Instanz für die Vergabe verantwortlich ist oder ob die Wahl einer Gruppenadresse dem Teilnehmer überlassen wird.

Skalierbarkeit: Das zentrale Problem der Gruppenkommunikation stellt die Skalierbarkeit dar, bei der mehrere Aspekte beachtet werden müssen:

Gruppengröße: Dieser Faktor bezieht sich auf die Anzahl der Teilnehmer einer Gruppe. Große Gruppen können dabei mehrere hundert oder tausend Teilnehmer umfassen, was z. B. bei verteilten Simulationen oder verteilten Spielen vorkommen kann. Ein mit der Gruppengröße einhergehendes Problem ist die Dynamik der Gruppe, also die Fluktuation der Teilnehmer innerhalb der Gruppe. Das stellt insbesondere hohe Anforderungen an die Gruppenverwaltung und an die dazugehörige Signalisierung.

Bekanntheit innerhalb der Gruppe: Hierunter wird der Grad der Kenntnis der Teilnehmer einer Gruppe verstanden. Die Bekanntheit ist dabei vor allem für eine zuverlässige Kommunikation wichtig. Hat die Gruppe eine hohe Dynamik, kann es schwierig sein, alle Gruppenwechsel bei allen Teilnehmern bekannt zu geben. Bei anonymen Gruppen muss die Identität der Teilnehmer nicht bekannt sein.

Topologie der Gruppe: Das ist die geografische Verteilung einer Gruppe. Die Topologie hat Einfluss auf die Laufzeit der Daten und besonders auf die Variation der Laufzeiten bei verschiedenen Gruppenteilnehmern.

Heterogenität: Zur Heterogenität der Teilnehmer werden die unterschiedlichen Anbindungsmöglichkeiten (z. B. Hochleistungsnetz, lokales Netz, drahtloses Netz oder ISDN) und auch die Endsysteme selbst verstanden. Diese Heterogenität hat wiederum Einfluss auf die Qualität der Datenübertragung.

2.1.3. IP-Multicast

IP-Multicast stellt eine sehr weit verbreitete Technik dar, die Gruppenkommunikation in IP-basierten Netzen unterstützt. IP ermöglicht eine Multicast-Kommunikation durch die Verwendung von Klasse D Adressen (auch als Multicast-Adressen bezeichnet). Jede Multicast-Adresse identifiziert eine Gruppe. Sendet ein Rechner ein Paket an ein Multicast-Adresse, wird es an alle Mitglieder der adressierten Gruppe zugestellt. Wie bei IP-Unicast gibt es keine Garantien, dass alle Gruppenmitglieder das Paket erhalten.

Häufig wird im Internet-Kontext statt Gruppenkommunikation der Begriff IP-Multicast verwendet. Streng genommen handelt es sich bei IP-Multicast um eine 1:n-Kommunikation, also eine eingeschränkte Form der Gruppenkommunikation. IP-Multicast ist ebenso wie IP-Unicast eine verbindungslose Kommunikation. Die Daten werden ohne Berücksichtigung der Senderadresse an die Gruppe der Empfänger weitergeleitet. Das macht IP-Multicast unabhängig vom jeweiligen Sender und es ermöglicht somit das gleichzeitige Senden von mehreren Datenquellen an die Empfängergruppe. IP-Multicast entspricht damit dem Paradigma der Multipeer-Kommunikationsform.

Es werden bei IP-Multicast zwei Typen von Adressen unterstützt: permanente und temporäre (siehe auch Unterkapitel 2.1.1, Lebensdauer einer Gruppe). Die permanenten Gruppenadressen sind z. B. für die Adressierung der Router (224.0.0.2) oder der Endsysteme (224.0.0.1) in einem Subnetz reserviert. Die temporären Adressen müssen vor der Nutzung erstellt werden. Hierzu fragt ein Rechner bei einem Beitritt oder Austritt zu einer Gruppe den lokalen Multicast-Router an.

Die Multicast-Router sind für das Multicast-Routing zwischen den Subnetzen verantwortlich. Beim Multicast-Routing ist eine ganz andere Ausgangssituation als beim herkömmlichen Punkt-zu-Punkt-Routing vorhanden, da jetzt eine Gruppe von Empfängern involviert ist. Zudem kann sich die Gruppenzusammensetzung im Laufe der Zeit ändern. Die Randbedingungen für das Multicast-Routing sind allerdings wie beim Unicast-Routing, das Routing sollte so effizient wie möglich sein, die Netzlast soll minimiert und Schleifen und Konzentrationspunkte vermieden werden. Es existieren eine

Reihe von Multicast-Routingprotokollen, die zum Teil gleiche Aufgaben erfüllen, zum Teil aber auch nur für bestimmte Gruppentypen geeignet sind. So existieren Protokolle für geografisch weit gestreute Gruppen (z. B. PIM Sparse Mode [13]) oder für Gruppen mit hoher Teilnehmerdichte (z. B. PIM Dense Mode [14]). Auf das umfangreiche Gebiet des Multicast-Routing wird im Folgenden aber nicht weiter eingegangen.

Für die Gruppenan- und abmeldungen der Endsysteme beim Multicast-Router wird ein Frage/Antwort-Protokoll namens IGMP (Internet Group Management Protocol [15]) benutzt. Das Protokoll dient der Gruppenverwaltung in einem Subnetz. IGMP ist genauso wie ICMP [16] integraler Bestandteil von IP. Die Multicast-Router benutzen IGMP, um Informationen über die Gruppenzugehörigkeit der angeschlossenen Systeme zu erhalten.

Die Einführung eines neuen Dienstes wie IP-Multicast erfordert im Allgemeinen, dass Endsysteme und Router mit dieser neuen Technologie ausgestattet werden müssen. Um hier eine schnelle Einführung zu ermöglichen, ist eine Übergangslösung gewählt worden. Es wurde ein Overlay-Netzwerk geschaffen, das sogenannte MBone (Multicast Backbone On the interNEt [17]). Das MBone besteht aus multicast-fähigen Teilnetzen und Verbindungen zwischen diesen Netzen. Die Verbindungen zwischen den Teilnetzen werden durch Tunnel realisiert. Diese Tunnel werden zwischen den Multicast-Routern etabliert und ermöglichen eine Überbrückung von Unicast-Routern, die nicht multicast-fähig sind.

Die Reichweite von IP-Multicast-Gruppen kann zudem noch begrenzt werden. Hierzu wird der TTL-Wert im IP-Paketkopf eingesetzt. Damit wird erreicht, dass die Pakete einer Gruppe nur innerhalb eines vorgegebenen Bereiches weitergeleitet werden und nicht darüber hinaus. Das hat den Vorteil, dass ein gewisser Grad an Privatheit gewährleistet werden kann, indem Rechner außerhalb des Bereiches der Gruppe nicht beitreten können. Darüber hinaus ermöglicht die Begrenzung der Reichweite eine Mehrfachnutzung von Multicast-Adressen, womit der begrenzte Adressraum effizienter genutzt werden kann. Die Grundvoraussetzung hierfür ist, dass sich die geografischen Netzbereiche der Gruppen nicht überschneiden dürfen.

2.2. Asynchronous Transfer Mode

Der **A**synchronous **T**ransfer **M**ode (ATM) ist eine verbreitete Technologie für Hochgeschwindigkeitsnetze [7, 8, 18]. Dabei weist ATM Ähnlichkeiten zu paketvermittelten Technologien wie X.25 und Frame Relay auf. ATM transportiert die Daten in diskreten Stücken konstanter Größe, den sogenannten ATM-Zellen. Die ATM-Zellen werden dabei logischen Verbindungen zugeordnet. Damit ermöglicht es ATM, die Zellen über eine physikalische Schnittstelle zu multiplexen.

Das ATM-Übertragungsverfahren ist auf eine hohe Übertragungsgeschwindigkeit hin optimiert und mit minimalen Fehler- und Flusskontrollfähigkeiten ausgestattet. Dadurch wird der Verarbeitungsaufwand für die ATM-Zellen und die Anzahl der Zusatzinformationen in den ATM-Zellen reduziert, was es ATM ermöglicht, hohe Datenraten ($> 1Gbit/s$) zu unterstützen. Darüber hinaus vereinfacht die Verwendung von Zel-

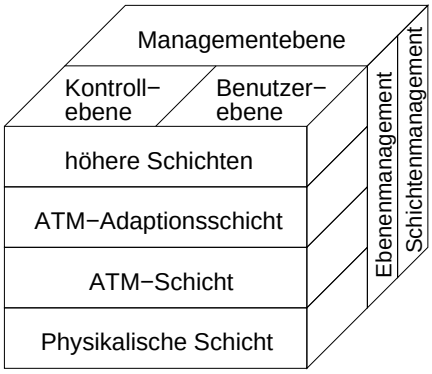


Abbildung 2.1.: ATM Protokollarchitektur.

len konstanter Größe die Verarbeitung in den ATM-Knoten und erlaubt eine effiziente Hardwareunterstützung bei der Weiterleitung der Zellen, was wiederum hohe Datenraten ermöglicht. Desweiteren unterstützt ATM verschiedene Diensttypen und -qualitäten, wobei die Einhaltung der vereinbarten Verkehrsparameter von ATM garantiert wird.

In diesem Unterkapitel werden die Schlüsselelemente der ATM-Technologie beschrieben, die für diese Arbeit relevant sind. Zuerst wird die Protokollarchitektur von ATM in Unterkapitel 2.2.1 vorgestellt, anschließend der Einsatz von logischen Verbindungen (Unterkapitel 2.2.2) und die ATM-Zellenstruktur (Unterkapitel 2.2.3). Darauf folgt eine Beschreibung der von ATM unterstützten Diensttypen in Unterkapitel 2.2.4 und die Beschreibung der ATM-Adaptionsschicht in Unterkapitel 2.2.5, mit der übergeordnete Protokolle auf ATM abgebildet werden. Am Ende wird die Signalisierung für den Verbindungsaufbau und das PNNI-Routing in Unterkapitel 2.2.6 kurz vorgestellt.

2.2.1. Protokollarchitektur

Die Protokollarchitektur von ATM basiert auf Standards der ITU-T [19] und ist in Abbildung 2.1 dargestellt. Die Abbildung 2.1 zeigt die Basisarchitektur der Schnittstelle zwischen Benutzer und Netzwerk. Die physikalische Schicht umfasst die Spezifikation des Übertragungsmediums und ein Signalkodierungsschema.

Zwei Schichten der Protokollarchitektur beziehen sich auf die ATM-Funktionalität. Es gibt eine ATM-Schicht, die von allen Diensten gemeinsam benutzt wird und eine Adaptionsschicht, die spezifisch für jeden Dienst ist. Die ATM-Schicht definiert die Übertragung von Zellen fester Größe und die Verwendung von logischen Verbindungen. Die Adaptionsschicht ist für die Abbildung von Informationen der höheren Schichten auf ATM-Zellen verantwortlich.

Zusätzlich hat das Referenzmodell aus Abbildung 2.1 drei getrennte Ebenen:

Benutzerebene: Zuständig für den Transport von Benutzerdaten und daran angeschlossene Kontrollen (z. B. Fluss- und Fehlerkontrolle).

Kontrollebene: Diese Ebene leistet die Rufkontrolle und beinhaltet Verbindungskontrollfunktionen.

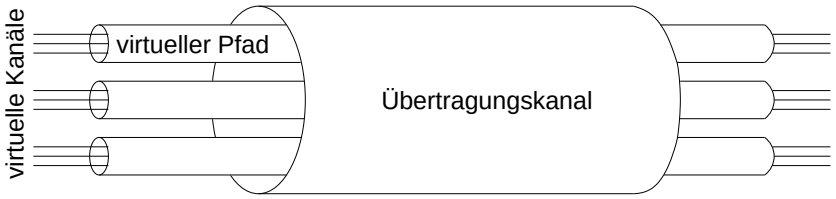


Abbildung 2.2.: Aufbau der logischen ATM Verbindungen.

Managementebene: Die Managementebene teilt sich auf in das Schichten- und das Ebenenmanagement. Das Ebenenmanagement führt Funktionen aus, die das System als Ganzes betreffen und es bietet Koordinationsfunktionen zwischen allen Ebenen an. Hinzu kommt ein Schichtenmanagement, das Funktionen für Ressourcen und Parameter zur Verfügung stellt, die zu den Protokollen gehören.

2.2.2. Logische Verbindungen bei ATM

Logische Verbindungen werden bei ATM als virtuelle Kanalverbindungen oder kurz VCC (VCC = Virtual Channel Connection) bezeichnet. Ein VCC stellt die Basiseinheit für die Zellenweiterleitung im ATM-Netzwerk dar. Der VCC wird zwischen zwei Endbenutzern im ATM-Netzwerk aufgebaut und bietet diesen einen vollduplex Datenfluss mit variabler Bitrate. VCCs werden des Weiteren für die Kontrollsignalisierung zwischen Benutzer und Netzwerk und für das Netzwerkmanagement und Routing innerhalb eines ATM-Netzwerkes eingesetzt.

Für ATM ist noch eine weitere logische Unterteilung eingeführt worden, die das Konzept von virtuellen Pfaden (VP = Virtual Path, Abbildung 2.2) aufgreift. Eine virtuelle Pfadverbindung (VPC = Virtual Path Connection) ist ein Bündel von virtuellen Kanälen, die dieselben Endpunkte haben. Damit können die Zellen aller VCCs innerhalb des VPCs über den gleichen Pfad geleitet werden.

Die Endpunkte eines VCC können Endbenutzer, Netzwerkeinheiten oder ein Endbenutzer und eine Netzwerkeinheit sein. In allen Fällen wird die Zellenreihenfolge beibehalten, also die Zellen werden in derselben Ordnung empfangen, in der sie gesendet worden sind. Zudem gibt es noch eine Reihe weiter Charakteristika, die für einen VCC bestimmt werden können:

Dienstgüte: Einer Verbindung kann eine Dienstgüte zugeordnet werden, die über Parameter wie Zellverlustrate oder Zellverzögerungsvarianz spezifiziert wird.

Verkehrsparameter: Für jede Verbindung können Verkehrsparameter zwischen Benutzer und Netzwerk ausgehandelt werden. Die Verkehrsparameter können hierbei z. B. die mittlere Rate oder Spitzenrate einer Verbindung sein. Das Netzwerk kontrolliert die Einhaltung der Verkehrsparameter für jede Verbindung, um garantieren zu können, dass der Verkehrsvertrag vom Benutzer eingehalten wird.

Permanente und geswitchte virtuelle Verbindungen: Eine geswitchte virtuelle Verbindung (SVC = Switched Virtual Circuit) wird nach Bedarf auf- und abgebaut.

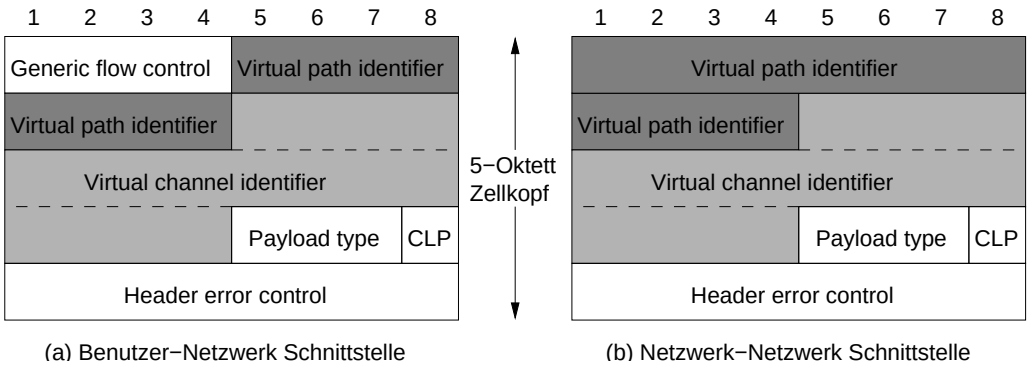


Abbildung 2.3.: Format des ATM-Zellenkopfes.

Hierfür existiert eine Verbindungskontrollsignalisierung, die für Auf- und Abbau der Verbindungen verantwortlich ist. Die permanenten virtuellen Verbindungen (PVC = Permanent Virtual Circuit) sind Langzeitverbindungen und werden über manuelle Konfiguration oder das Netzwerkmanagement eingerichtet.

2.2.3. ATM-Zellen

ATM verwendet Dateneinheiten fester Größe, sogenannte Zellen. Jede Zelle besteht aus einem 5 Oktett großen Kopf und einem 48 Oktett großen Informationsfeld. Zellen fester Größe können sehr effizient geswitched werden, was wichtig für die Unterstützung hoher Datenraten ist. Eine Implementierung der Switching-Mechanismen in Hardware wird hierdurch ebenfalls vereinfacht. Darüber hinaus kann die Warteschlangenverzögerung für höher priorisierte Zellen reduziert werden. Durch die geringe Zellengröße muss eine Zelle kürzer warten, wenn sie nach einer niedriger priorisierten Zelle ankommt, die gerade am Ausgang übertragen wird. Die geringe Zellengröße hat den inhärenten Nachteil des durch den Paketkopf bedingten Overheads, der $\sim$10% des Datenaufkommens auf einer ATM-Leitung ausmacht.

Die Abbildung 2.3 zeigt den ATM-Zellenkopf. Es existieren zwei ähnliche Zellenkopfformate, je nachdem ob die ATM-Zelle zwischen Benutzer und Netzwerk oder zwischen zwei Netzwerkeinheiten transportiert wird. Das Generic-flow-control-Feld hat nur eine lokale Funktionalität und wird im ATM-Netzwerk nicht behalten. Statt dessen wird im ATM-Netzwerk das Feld für den Virtual path identifier von 8 auf 12 Bit erweitert.

Der Virtual path identifier (VPI) stellt ein Label für das Routing im Netzwerk dar. Genauso wird der Virtual channel identifier (VCI) verwendet. Anhand von VPI und VCI wird die Zelle einer logischen Verbindung (VCC) zugeordnet. Das Payload-Type-Feld gibt den Typ der Information im Informationsfeld an. Es wird dabei zwischen Benutzerdaten-, Netzwerkmanagement- oder Ressourcenmanagement-Zelle unterschieden. Das Cell-Loss-Priority-Bit (CLP) wird für die Lenkung der Zelle in Stausituationen benutzt. Ein Wert von 0 markiert eine Zelle mit höherer Priorität, die nach Möglichkeit nicht verworfen werden soll. Ein Wert von 1 ist bei Zellen gesetzt, die verworfen werden können. Dieses Feld wird in der Regel bei Zellen gesetzt, die nicht dem ausgehandelten

Verkehrsvertrag entsprechen. Das Header-Error-Control-Feld enthält eine Prüfsumme über die vorhergegangenen 32 Bit.

2.2.4. ATM Dienstkategorien

Die ATM-Technologie ist entworfen worden, um viele verschiedene Verkehrstypen gleichzeitig transportieren zu können. Hierzu gehören Realzeitdatenströme, wie Sprache und Video, und der Transport von voluminösen Bulk-Daten, wie bei FTP. Jeder Datenstrom wird als ein Zellenstrom innerhalb einer logischen Verbindung aufgefasst. Die Art und Weise, in der jeder Datenstrom im Netzwerk behandelt wird, hängt dabei von den Charakteristika des Datenstroms und den Anforderungen der Anwendungen ab, z. B. muss ein Realzeitvideodatenstrom mit einem Minimum an Verzögerungsschwankungen ausgeliefert werden. Das ATM Forum hat hierzu fünf Dienste in zwei Kategorien definiert, die ein Endsystem benutzen kann, um seinen benötigten Dienst zu bestimmen:

Realzeit-Dienste: Die wichtigste Unterscheidung zwischen Anwendungen betrifft den Umfang an Verzögerung und Verzögerungsschwankungen (Jitter), den Anwendungen akzeptieren können. Realzeit-Dienste umfassen typischerweise einen Informationsfluss, der eine unmittelbare Verarbeitung der Daten beim Empfänger erfordert. Zum Beispiel erwartet ein Benutzer, dass ein Strom von Audio- oder Videodaten in einem kontinuierlichen Fluss dargestellt werden kann, Unterbrechungen stellen einen Verlust der Übertragungsqualität dar.

Constant Bit Rate (CBR): Diese Dienstkategorie stellt eine feste, kontinuierlich vorhandene Datenrate zwischen den Endsystemen zur Verfügung. Anwendungsbeispiele für CBR sind Videokonferenzen, Sprachdaten (Telefon) oder Audio/Video Datenverteilung.

Real-time Variable Bit Rate (rt-VBR): Diese Kategorie ist definiert für verzögerungssensitive Anwendungen, die strenge Anforderungen an Verzögerung und Verzögerungsschwankung stellen. Der Unterschied zu CBR ist, dass die Datenrate schwanken kann, wie z. B. bei komprimiertem Video. Der Dienst rt-VBR erlaubt dem Netzwerk eine höhere Flexibilität als bei CBR. Das Netzwerk kann die Verbindungen statistisch multiplexen und weniger Ressourcen vergeben, was aber mit einer etwas höheren Paketverlustwahrscheinlichkeit verbunden ist.

Nicht-Realzeit-Dienste: Die Dienste dieser Kategorie sind spezifiziert für Anwendungen, die eine burstartige Verkehrscharakteristik, aber keine strengen Begrenzungen bezüglich Verzögerung und Verzögerungsschwankung haben. Das Netzwerk hat bei diesen Diensten eine größere Flexibilität in der Behandlung des Datenstroms und kann einen größeren Nutzen aus dem statistischen Multiplexen und der Pufferung von ATM-Zellen ziehen, um die Netzwerkeffizienz zu erhöhen.

Non-real-time Variable Bit Rate (nrt-VBR): Bei einigen Nicht-Realzeit-Anwendungen ist es möglich, das zu erwartende Verkehrsaufkommen im Vorhinein

zu charakterisieren, so dass das ATM-Netz hierfür eine bessere Dienstqualität in Bezug auf Verlust und Verzögerung anbieten kann. Diese Anwendungen können den Dienst nrt-VBR benutzen. Die Anwendung kann die mittlere Rate, die Spitzenrate und die 'Burstiness' angeben. Anwendungsbeispiele sind Dienste mit kritischen Antwortzeiten, z. B. Prozessüberwachung, Banken-Transaktionen oder Sitzplatzreservierungen.

Unspecified Bit Rate (UBR): Diese Kategorie entspricht dem Best-Effort-Dienst bei IP. Es werden keinerlei Garantien bzgl. Verlust und Verzögerung gemacht. UBR nutzt die Ressourcen vom Netzwerk, die von den höherwertigen Diensten (CBR, rt-VBR und nrt-VBR) nicht benötigt werden. Daher ist die Übertragungsqualität und -kapazität bei UBR Schwankungen unterworfen. Mit ABR können diese Schwankungen reduziert werden.

Available Bit Rate (ABR): Anwendungen mit burstartigen Verkehr, die ein zuverlässiges Übertragungsprotokoll wie TCP verwenden, können Staus im Netz erkennen und darauf reagieren. Aber TCP kennt keine Mechanismen, um die Ressourcen im Netzwerk fair zwischen den Verbindungen aufzuteilen. Des Weiteren kann TCP Staus nicht so effizient vermeiden, wie es durch explizite Stauinformationen im Netzwerk möglich wäre. Beim ABR-Dienst versucht das Netzwerk, die noch vorhandenen Ressourcen gleichmäßig auf die existierenden ABR-Verbindungen aufzuteilen und die Verkehrsmenge, die ins Netzwerk gelangt, gleich beim Endsystem(Benutzer) zu regulieren. Bei ABR kann zudem von den Anwendungen eine minimale Datenrate reserviert werden, die in jedem Fall bei diesem Dienst garantiert wird.

2.2.5. Die ATM-Adaptionsschicht

Eine wesentliche Aufgabe der Adaptionsschicht (siehe auch Abbildung 2.1) besteht in der Segmentierung und Reassemblierung der Daten höherer Schichten in ATM-Zellen. Des Weiteren kann die Behandlung von Übertragungsfehlern, Zellverlusten und Flusskontrolle zu den Aufgaben der Adaptionsschicht gehören. Um die Anzahl der verschiedenen AAL-Protokolle und -Dienste zu minimieren, hat die ITU-T vier Dienstklassen definiert, die eine große Menge von Anforderungen umfassen (Tabelle 2.1). Es wurden ursprünglich vier AAL-Typen definiert: AAL1 für Dienste mit konstanten Bitraten, AAL2 für Dienste mit variablen Bitraten, AAL3 und AAL4 wurden zu AAL3/4 zusammengefasst und sind für die Übertragung von Daten ohne Realzeit- und Übertragungsratenanforderungen gedacht. Nachträglich ist noch eine fünfte, stark vereinfachte Version von AAL3/4 spezifiziert worden: AAL5. Heutzutage wird praktisch nur noch AAL5 für die Übertragung von IP-Datenpaketen verwendet.

Das Protokoll AAL5 wurde von der ITU-T eingeführt, um einen einfachen Datentransport für höhere Schichten zur Verfügung zu stellen, der nur einen geringen zusätzlichen Overhead bei der Übertragung und im Protokoll hat. Im Vergleich zu AAL3/4 ist der Overhead wesentlich geringer, da das gesamte Datenfeld einer Zelle von AAL5 verwendet wird.

	Class A	Class B	Class C	Class D
Zeit-Kompensation	erforderlich		nicht erforderlich	
Bit Rate	konstant	variabel		
Verbindungsmodus	verbindungsorientiert			verbindungslos
AAL Protokoll	Typ 1	Typ 2	Typ 3, Typ 5	Typ 4

Tabelle 2.1.: AAL Dienstklassifikationen

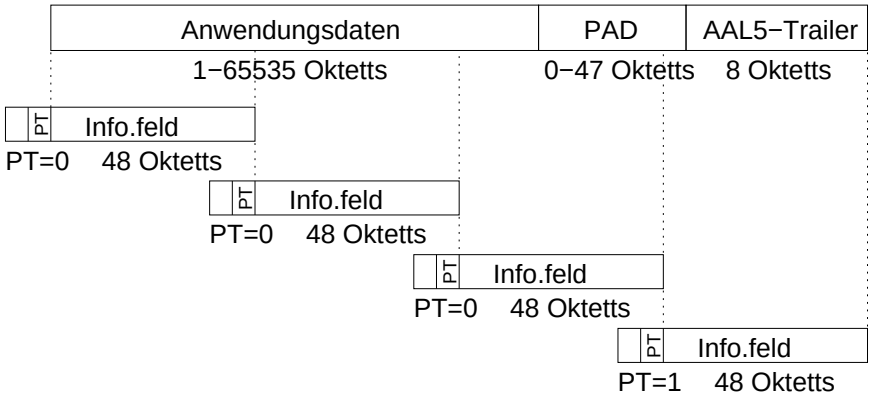


Abbildung 2.4.: Übertragung eines AAL5 Datenpaketes.

Das Anwendungsdatenpaket kann zwischen 1 und 65535 Oktetts groß sein (Abbildung 2.4). An das Datenpaket wird ein 8 Oktett großer Trailer angehängt und dazwischen wird mit einem variablen PAD-Feld auf ein ganzes Vielfaches von 48 Oktett aufgefüllt. Diese Daten werden jetzt in Zellen segmentiert, wobei das Payload-Type-Feld benutzt wird, um das Paketende zu markieren. Jeweils bei der letzten Zelle eines AAL5-Paketes hat der Payload Type den Wert 1. Die acht Oktett des AAL5-Trailers setzen sich wie folgt zusammen:

User-to-user indication (1 Oktett): Für die Übertragung von Benutzerinformationen.

Common part indicator (1 Oktett): Definiert die Interpretation der folgenden zwei Felder, momentan ist nur eine einzige, nämlich die hier dargestellte, Interpretation definiert.

Length (2 Oktett): Die Länge des Anwendungsdatenpaketes.

CRC (4 Oktett): Prüfsumme über das gesamte AAL5-Paket inklusive Anwendungsdaten.

Die Reduzierung des Overheads bei AAL5 hat verschiedene Folgen, die bei der Verwendung von AAL5 beachtet werden müssen:

- Da keine Sequenznummern in den Zellen existieren, muss der Empfänger voraussetzen, dass alle Zellen eines Paketes in der korrekten Reihenfolge ankommen. Das kann mit der Prüfsumme im Trailer kontrolliert werden.

- Es ist nicht möglich, mehrere AAL5-Pakete einer Verbindung ineinander zu verschachteln. Die AAL5-Pakete müssen sequenziell gesendet werden.

- Der Verlust einer Zelle macht das ganze AAL5-Paket unbrauchbar. Das kann aber erst festgestellt werden, nachdem das AAL5-Paket zusammengefügt und die Längeninformation aus dem Trailer mit der tatsächlichen Länge verglichen worden ist. Dieses Problem ist aber auch bei AAL3/4 vorhanden.

2.2.6. Signalisierung und Routing in ATM-Netzen

Für die Etablierung einer geswitchten logischen Verbindung (SVC) zwischen zwei Endsystemen wird ein Signalisierungsprotokoll benötigt, das den Weg im Netzwerk festlegt und auf diesem Weg die notwendigen logischen und physikalischen Ressourcen reserviert. Für diese Wegewahl im ATM-Netz ist das PNNI-Routingprotokoll [20] zuständig und für den darauf basierenden Verbindungsauf- und -abbau die UNI-Signalisierung [21]. Zunächst wird anhand von Weg-Zeit-Diagrammen die Signalisierung für den Auf- und Abbau von Punkt-zu-Punkt- und Punkt-zu-Mehrpunkt-Verbindungen vorgestellt. Anschließend folgt ein Überblick über das hierarchisch strukturierte PNNI-Routingprotokoll.

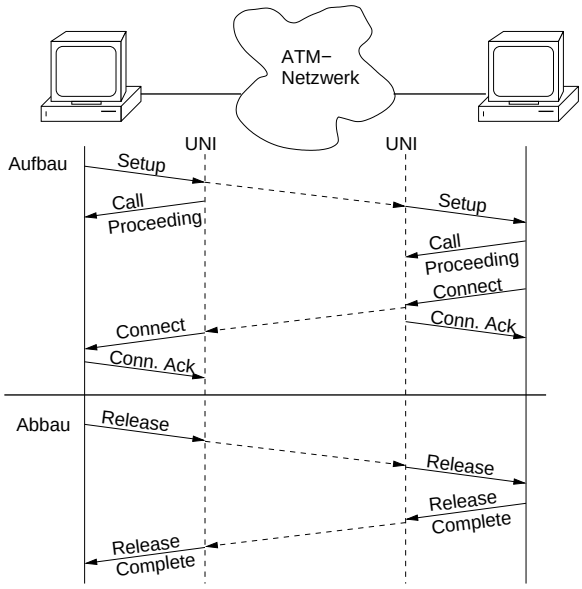
UNI-Signalisierung

Die Abbildung 2.5(a) zeigt den Auf- und Abbau von ATM-Verbindungen [22]. Der Verbindungsaufbau wird immer von der Quelle initiiert, der Abbau kann von der Quelle oder der Senke eingeleitet werden. In der Abbildung 2.5(a) ist nur der Verbindungsabbau von der Quelle ausgehend dargestellt. Für die Weiterleitung in der ATM-Wolke ist das PNNI-Protokoll zuständig, das aber die gleichen Nachrichtenformate wie die UNI-Signalisierung verwendet.

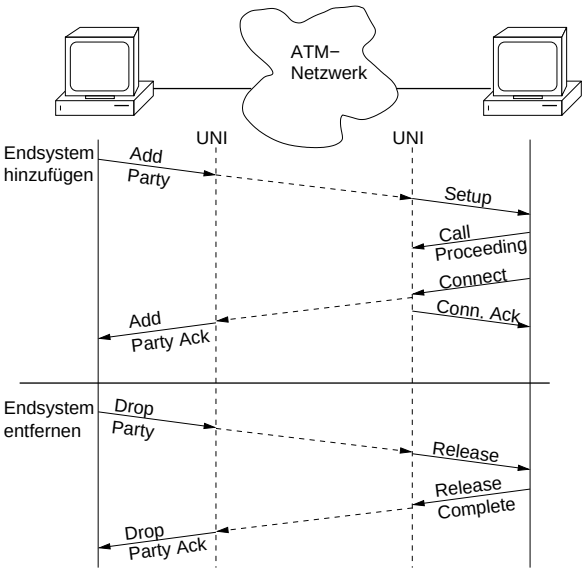
Ist eine Punkt-zu-Punkt-Verbindung aufgebaut, können dieser Verbindung weitere Endsysteme als Empfänger hinzugefügt werden (Abbildung 2.5(b)). Diese Möglichkeit muss bereits beim Verbindungsaufbau berücksichtigt werden. Ein wichtiger Unterschied zwischen Punkt-zu-Punkt- und Punkt-zu-Mehrpunkt-Verbindungen ist der unidirektionale Datentransport bei Mehrpunkt-Verbindungen. Bei Punkt-zu-Punkt-Verbindungen ist sowohl uni- als auch bidirektionaler Datentransport möglich.

Wie in Abbildung 2.5(b) zu sehen ist, besteht für die Senke keinerlei Unterschied zu einer Verbindungsannahme von einer Punkt-zu-Punkt-Verbindung. Nur bei der Quelle wird das Hinzufügen und Entfernen eines zusätzlichen Teilnehmers gesondert signalisiert.

Bei der aktuellen UNI-Version 4.0 kann nicht nur die Quelle weitere Empfänger zu einer Verbindung hinzufügen. Auch eine Senke kann einer Verbindung beitreten (Leaf Initiated Join), ohne dass dies von der Quelle eingeleitet wird. Darüber hinaus definiert UNI 4.0 Anycast für ATM-Netzwerke. Mittels Anycast können bekannte Server



(a) Aufbau und Abbau einer ATM-Verbindung.



(b) Hinzufügen und Entfernen eines Endsystems zu einer ATM-Verbindung.

Abbildung 2.5.: UNI Signalisierung.

und Dienste (z. B. bei LAN Emulation) adressiert werden. Hierzu sind zu den Unicast-Adressen in ATM spezielle Gruppenadressen hinzugekommen. Des Weiteren kann zu jeder Gruppenadresse eine Reichweite (Scope) festgelegt werden, die den Bereich des Routings für diese Gruppenadresse begrenzt.

PNNI-Routing

Das PNNI-Routingprotokoll ist für den Verbindungsaufbau im ATM-Netzwerk zuständig. Merkmale von PNNI sind eine hierarchische Unterteilung des Netzwerkes, Berücksichtigung der Dienstgüte beim Routing und quellen-basiertes Routing.

Das gesamte Netzwerk wird bei PNNI logisch in sogenannte Peer Groups strukturiert, welche wiederum zu übergeordneten Peer Groups zusammengefasst werden. Hierdurch entsteht ein logischer Baum, die PNNI-Hierarchie. Die Blätter des Baumes sind ATM Schalteinheiten. Endsysteme kommen in der Hierarchie nicht vor, es wird davon ausgegangen, dass die Endsysteme sternförmig an den ATM Schalteinheiten angeschlossen sind. Eine Peer Group ist eine Verwaltungseinheit und unabhängig von der physikalischen Netzstruktur. Durch die hierarchische Unterteilung der Peer Groups wird vermieden, dass jeder Knoten die komplette Struktur des Netzwerkes kennen muss.

Abbildung 2.6 zeigt ein Beispielnetzwerk, das in Hierarchieebenen logisch untergliedert ist. In der untersten Ebene besteht das Netz aus einzelnen Knoten, z. B. 1.1.2, 2.2.4, u.s.w., den ATM Schalteinheiten. Diese werden zu Peer Groups zusammengefasst, z. B. besteht die Peer Group 2.2 aus den Knoten 2.2.1, 2.2.2, 2.2.3 und 2.2.4. Ein ausgezeich-

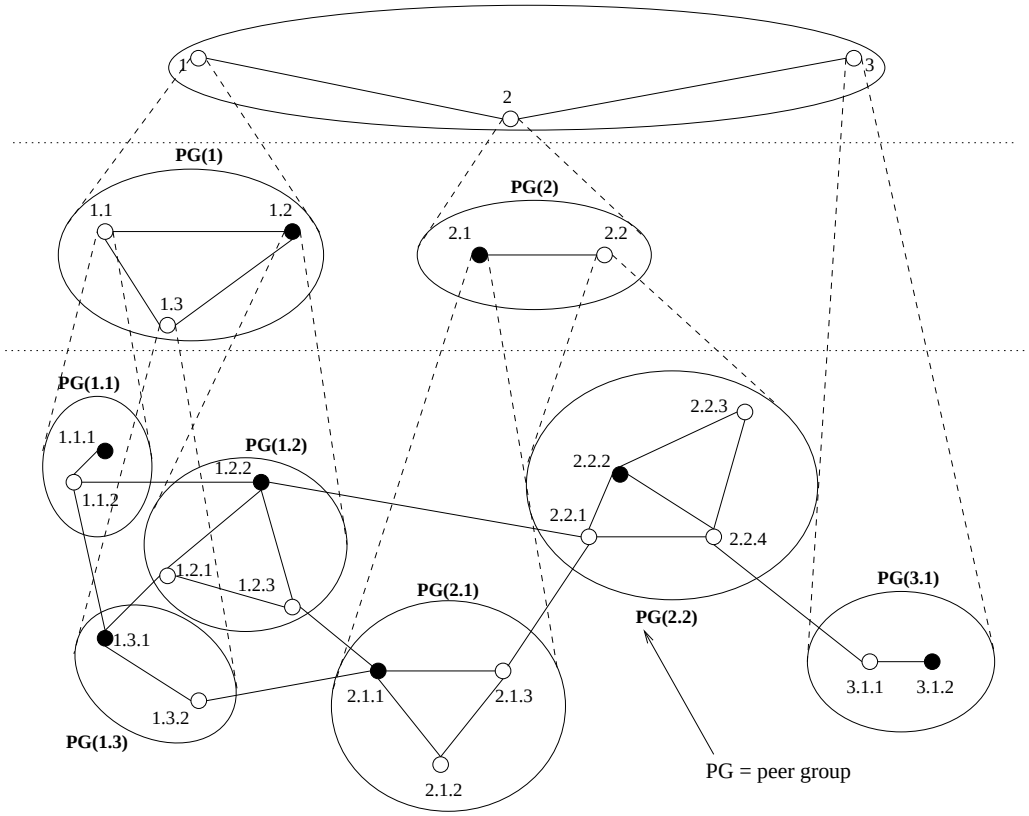


Abbildung 2.6.: Beispiel für eine PNNI-Hierarchie.

neter Knoten, der Peer Group Leader (in Abbildung 2.6 als ausgefüllter Kreis markiert), übernimmt die Aufgabe, die Peer Group, in welcher der Knoten selber enthalten ist, auf einer höheren Ebene zu repräsentieren. Diese Peer Group Leader werden dann wiederum zu einer Peer Group zusammengefasst (z. B. 2.1 und 2.2 zu 2), u.s.w. Welche Knoten zu einer Peer Group gehören, wird anhand der Adressen entschieden (falls nicht anders konfiguriert). Eine Peer Group besteht aus Knoten, die denselben Adressen-Präfix haben. Die Struktur der PNNI-Hierarchie wird somit durch die Adressvergabe bestimmt.

Innerhalb einer Peer Group haben alle Knoten dieselben Informationen bzgl. des Netzwerkes. Diese Informationen werden periodisch oder beim Eintreten einer Änderung ausgetauscht. Innerhalb einer Peer Group werden die Informationen über einen Broadcast-Mechanismus verteilt. Zwischen den Peer Groups sorgen die Peer Group Leader für einen Informationsaustausch zwischen den Ebenen in beiden Richtungen, also von der unteren zur höheren Ebene und umgekehrt.

Die Sicht eines Knotens auf das Netzwerk ist in Abbildung 2.7 dargestellt. Die Knoten 2.2.1, 2.2.2, 2.2.3, 2.2.4 haben dieselbe Sicht auf das Netzwerk, ihre Daten werden untereinander abgeglichen. Die Knoten kennen die Topologie ihrer eigenen Peer Group und die Topologien aller höher liegenden Peer Groups in denen die Knoten enthalten sind. Zusätzlich werden Informationen zu benachbarten Peer Groups (dargestellt durch die gepunkteten Linien) gespeichert. Diese benachbarten Peer Groups werden durch den

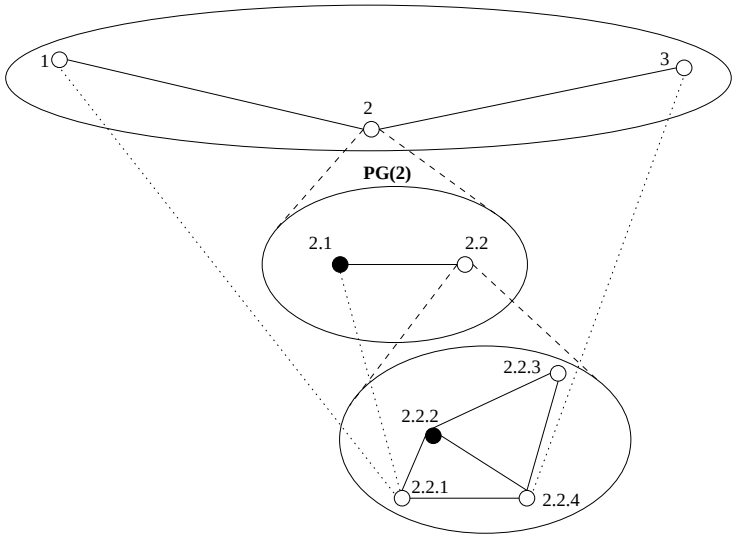


Abbildung 2.7.: Sicht eines Knotens auf das Netzwerk.

jeweils in der Hierarchie zuerst erreichbaren Peer Group Leader repräsentiert. Anhand dieser Informationen legt die Quelle die Route zur Senke fest. Dieser Weg wird an die höheren Hierarchieebene weitergegeben. Dort muss er noch weiter aufgelöst werden, so dass die reale Route bestimmt werden kann. Möchte z. B. der Knoten 2.2.1 eine Verbindung zum Knoten 3.1.2 aufbauen, so legt Knoten 2.2.1 die Route $2.2.1 \rightarrow 2.2 \rightarrow 2 \rightarrow 3$ fest, die dann in den höheren Knoten weiter aufgelöst werden muss, so dass am Ende die (mögliche) Route $2.2.1 \rightarrow 2.2.4 \rightarrow 3.1.1 \rightarrow 3.1.2$ gewählt wird.

Bei der Routingentscheidung wird auch versucht, die Dienstgüte zu berücksichtigen. Hierzu wird in den Hierarchieebenen eine abstrakte Repräsentation des Netzes angewendet, in der die Informationen über vorhandene Ressourcen akkumuliert werden.

2.3. IP über ATM

Die hohe Verbreitung von TCP/IP im Internet ist unbestritten. Durch die Ausbreitung des Internet und durch heterogene Netzstrukturen hat sich diese Protokollfamilie durchgesetzt. Für eine Etablierung von ATM ist es daher unumgänglich, dass ATM diese Protokollfamilie unterstützt. Ein Interworking zwischen IP und ATM ist notwendig.

ATM ist verbindungsorientiert aufgebaut und besitzt eine eigene Adressstruktur und eigene Routing-Funktionen. Mit ATM können Punkt-zu-Punkt- und Punkt-zu-Mehrpunkt-Verbindungen aufgebaut werden. Das Internet Protokoll (IP) ist hingegen verbindungslos organisiert. Die IP-Datenpakete werden auf Hop-by-Hop-Basis weitergeleitet, unabhängig vom darunter liegenden Netzwerk. IP ist darauf ausgelegt, im Zugangsbereich auf broadcast-fähigen Netzwerken zu arbeiten. Der Broadcast-Mechanismus wird z. B. zur Adressauflösung verwendet (ARP [12]).

Um die zwei unterschiedlichen Welten von ATM und IP miteinander zu verbinden, sind verschiedene Verfahren entwickelt worden. Im Folgenden soll nur die Variante

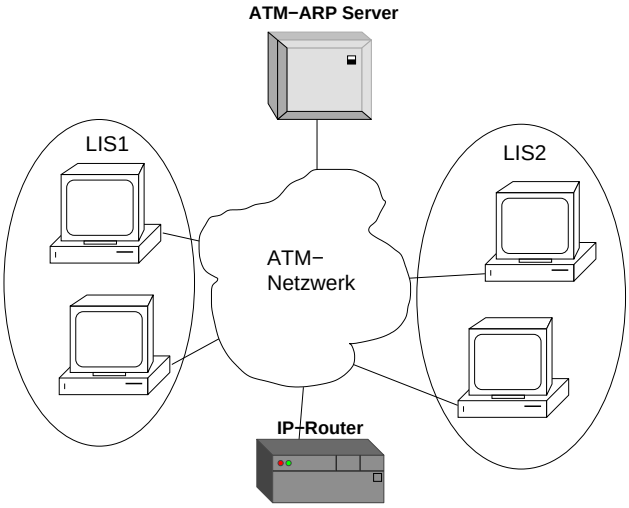


Abbildung 2.8.: Logische IP-Subnetze in einem ATM-Netz.

Classical-IP over ATM (CLIP) [23] vorgestellt werden. Die beiden anderen Möglichkeiten, LAN-Emulation (LANE [24]) und Multi-Protocol-over-ATM (MPOA [25]) werden nicht behandelt.

CLIP ist bereits erstmalig 1993 von der IETF veröffentlicht worden. Für die Übertragung von IP-Paketen über ATM müssen die IP-Pakete eingekapselt werden. Das geschieht mit der LLC/SNAP-Einkapselung [26], wodurch sich alle Pakete identifizieren lassen. Eine weitere Auswirkung der LLC/SNAP-Einkapselung ist, dass alle Verbindungen an der LLC-Schicht im Endsystem enden, und hier keine Änderungen in der IP-Vermittlungsschicht notwendig sind.

Für die Übertragung von IP-Paketen über ATM ist es aber hauptsächlich notwendig, die IP-Adressen in ATM-Adressen umzusetzen. Hierzu legt CLIP logische IP-Subnetze (LIS) auf ATM-Netzen fest. Das ist in Abbildung 2.8 dargestellt. Ein LIS besteht aus mehreren Endsystemen und Routern und einem ATMARP-Server. Dieser ATMARP-Server kann auch mehrere LIS bedienen, seine Aufgabe ist ausschließlich die Adressauflösung.

Jedes Endsystem ist mit der ATM-Adresse des ATMARP-Servers konfiguriert und baut eine Verbindung zu diesem Server auf. Das ATMARP-Protokoll ersetzt dabei das ARP-Protokoll von IP. Jedes Endsystem wird in periodischen Abständen (Standard: alle 20 Minuten) vom ATMARP-Server nach seiner IP-Adresse gefragt (inverse ATMARP-Meldung). Das ist in Abbildung 2.9(a) dargestellt. Möchte ein Endsystem ein Paket an eine IP-Adresse senden, so wird eine Anfrage vom Endsystem an den ATMARP-Server gestellt (Abbildung 2.9(b)), die mit der zugehörigen ATM-Adresse positiv oder ohne ATM-Adresse negativ beantwortet wird.

Hat das Endsystem die zugehörige ATM-Adresse erhalten, so kann es eine ATM-Verbindung zur Zieladresse aufbauen und die IP-Pakete auf dieser Verbindung zum anderen Endsystem übertragen. Die ATM-Adressen werden in den Endsystemen temporär in einem Cache gespeichert. Alle geöffneten ATM-Verbindungen werden nach einem Zeit-

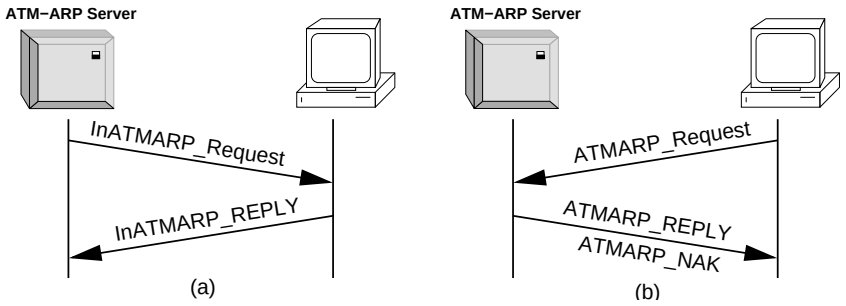


Abbildung 2.9.: ATMARP-Kommunikation: (a) Anfordern der Endsystem-IP-Adresse vom ATMARP-Server und (b) Abfrage einer ATM-Adresse vom Endsystem.

raum von 5 Minuten wieder geschlossen, wenn in diesem Zeitraum keine Pakete über die ATM-Verbindungen übertragen worden sind.

2.4. IP-Multicast über ATM

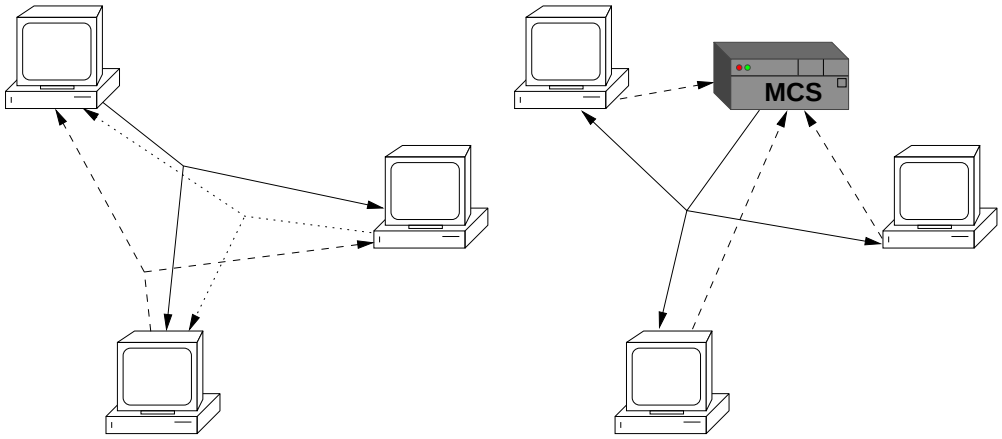
2.4.1. Multipeer-Emulation über ATM

Die ATM Technologie bietet eine rudimentäre Multicast-Unterstützung durch Punkt-zu-Mehrpunkt-Verbindungen an. Das stellt allerdings keine befriedigende Lösung dar, da diese Unterstützung nicht in der Lage ist, die Anforderungen der meisten Gruppenanwendungen in punkto Flexibilität und Skalierbarkeit zu erfüllen. In den aktuellen ATM-Standards existiert auch keine Abstraktion durch Gruppenadressen, die Sender müssen über die beteiligten Empfänger durch zusätzliche Mechanismen informiert werden. Darüber hinaus existiert keinerlei Unterstützung für eine Multipeer-Kommunikation, wie sie von IP-Multicast angeboten wird.

Es existieren zwei grundlegende Modelle für ATM, um die Multipeer-Kommunikation mit Punkt-zu-Mehrpunkt-Verbindungen emulieren zu können [27]:

VC Mesh: Jeder Sender unterhält pro Gruppe eine Punkt-zu-Mehrpunkt-Verbindung zu allen Empfängern der Gruppe. Wenn ein Empfänger der Gruppe beitritt oder die Gruppe verlässt, müssen die Verbindungen aller Sender aktualisiert werden. Ferner muss ein Empfänger für jeden aktiven Sender eine ATM-Verbindung terminieren. Das Prinzip des VC Mesh ist in Abbildung 2.10(a) dargestellt.

Multicast Server (MCS): In diesem Modell wird ein Server (der MCS) in einem Cluster ausgewählt, an den jeder Sender seine Daten überträgt. Der MCS etabliert eine Punkt-zu-Mehrpunkt-Verbindung zu allen Empfängern und leitet die ankommenden Datenpakete auf dieser Verbindung weiter (Abbildung 2.10(b)). Der MCS stellt eine Art Proxy-Server dar, der alle eingehenden Verbindungen zusammenfasst. Ein Nebeneffekt hierbei ist, dass einige Teilnehmer reflektierte Datenpakete



(a) VC-Mesh-Schema. (b) MCS-Schema.

Abbildung 2.10.: Grundlegende Modelle für die Multipeer-Kommunikation über ATM (alle Teilnehmer sind gleichzeitig Sender und Empfänger).

erhalten: Sender, die gleichzeitig Empfänger in der Gruppe sind, bekommen eine Kopie ihrer gesendeten Datenpakete zurück. Diese reflektierten Datenpakete müssen von der Quelle wieder ausgesondert werden.

Beide Modelle haben Vor- und Nachteile. In Bezug auf den Datendurchsatz ist die VC Mesh Lösung vorzuziehen, da hier kein Verkehrskonzentrationspunkt, wie der MCS, vorhanden ist. Die Verzögerungen bei der Datenübertragung mit VC Mesh sind ebenfalls geringer als beim MCS-Schema, da hier die Reassemblierung der Datenpakete beim MCS vermieden wird. Jedoch ist das MCS-Schema besser geeignet für dynamische Empfängergruppen, da hier Sender und Empfänger durch den MCS getrennt sind und somit eine bessere Gruppenteilnehmerkontrolle im MCS möglich ist. Bezüglich des Ressourcenverbrauchs hat das MCS-Schema ebenfalls Vorteile gegenüber dem VC-Mesh-Schema, da beim MCS-Schema nur zwei ATM-Verbindungen pro Endsystem nötig sind.

Ein inhärentes Problem des MCS-Schemas sind reflektierte Datenpaketen. Ist ein Gruppenteilnehmer sowohl Sender als auch Empfänger in der Gruppe, so bekommt er seine eigenen Datenpakete vom MCS nochmals zugesendet. Dieses Verhalten wird als Paketreflexion bezeichnet und kann zu Problemen in den darüber angesiedelten Anwendungen führen. Die Anwendungen sind in der Regel an IP Multicast angepasst, wo keine Paketreflexionen auftreten. Daher ist es notwendig, reflektierte Datenpakete herauszufiltern und nicht an höhere Schichten weiterzuleiten. Darüber hinaus führen reflektierte Datenpakete zu einer höheren Netzwerkbelastung.

2.4.2. MARS

Trotz der beiden Schemata zur Abbildung einer Multipeer-Kommunikation auf ATM-Verbindungen, besteht immer noch das Problem, wie eine Gruppenadresse einer höheren

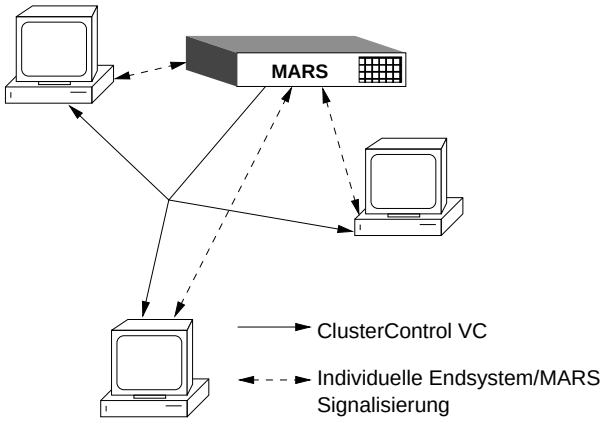


Abbildung 2.11.: ATM-Verbindungen für die Signalisierung zwischen MARS und Endsystemen.

Schicht auf ATM abgebildet werden kann. Eine Lösung hierfür ist der sogenannte MARS (**M**ulticast **A**ddress **R**esolution **S**erver [28]). Dieser Ansatz ist von der IETF als Standard für die IP-Multicast Emulation über ATM-Netzen vorgeschlagen worden. Der MARS basiert auf einer Erweiterung des ATMARP-Servers (siehe Unterkapitel 2.3). Der MARS unterhält erweiterte Tabellen mit Abbildungen von jeweils einer IP-Multicast-Adresse auf eine Menge von ATM-Adressen, die die angemeldeten Empfänger der Gruppe repräsentieren.

Ein MARS ist jeweils für einen MARS-Cluster verantwortlich. Ein MARS-Cluster ist im Allgemeinen identisch mit einem LIS (**L**ogical **I**P **S**ubnetz, siehe Unterkapitel 2.3). Diese Voraussetzung ist nicht zwingend notwendig, aber aus praktischen Erwägungen sinnvoll, da dann die zu administrierenden Bereiche identisch sind. MARS-Cluster werden untereinander über IP-Multicast-Router verbunden, die auch die Verbindung zu einem nicht-ATM Subnetz herstellen. Innerhalb eines MARS-Clusters stellt der MARS die zentrale Registrierung für IP-Multicast Gruppenadressen dar. Für die Verwaltung und Signalisierung der Gruppenmitgliedschaften und deren Änderungen benötigt der MARS und die beteiligten Endsysteme zusätzliche Signalisierungsverbindungen und ein Signalisierungsprotokoll.

Signalisierung zwischen MARS und Endsystem

Für die Signalisierung unterhält jedes Endsystem eine bidirektionale ATM-Verbindung zum MARS (siehe Abbildung 2.11). Zusätzlich unterhält der MARS eine Punkt-zu-Mehrpunkt-Verbindung zu allen Endsystemen, den **C**luster**C**ontrol **VC** (CCVC). Über diese Verbindungen können Endsysteme ihren Gruppenbeitritt oder -austritt signalisieren, unabhängig vom gewählten Verteilungsschema (VC Mesh oder MCS). Die Signalisierung hierzu soll an dem Beispiel in Abbildung 2.12 verdeutlicht werden. Gezeigt ist in Abbildung 2.12 die Signalisierung für den Beitritt von Endsystem 1 als Empfänger und Sender zu einer Gruppe und den anschließenden Austritt von Endsystem 1. Das Endsystem 1 meldet sich bei der Gruppe als Empfänger mit der Gruppenadresse (gid) und

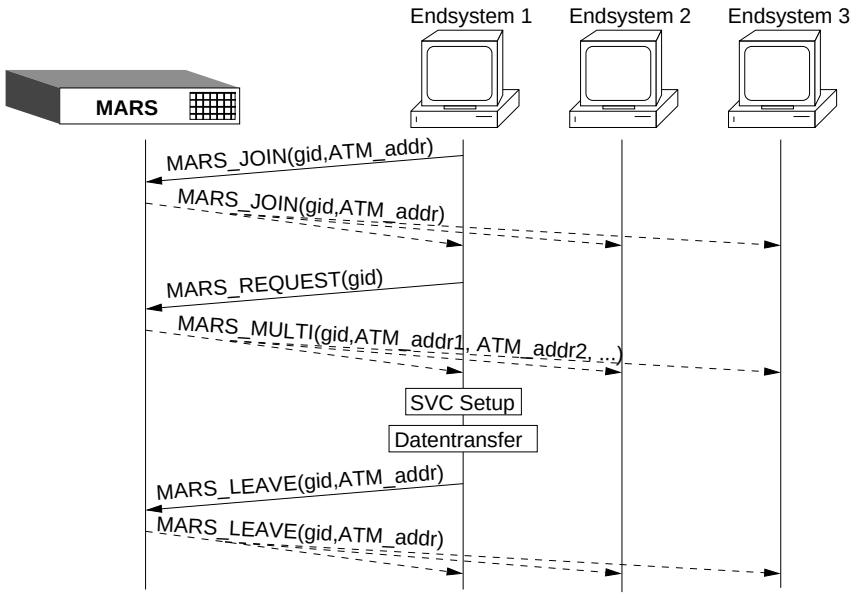


Abbildung 2.12.: Multicast Adressauflösung mit MARS.

der eigenen ATM-Adresse an (MARS_JOIN). Der MARS verteilt diese Anmeldung über den ClusterControl VC an alle angeschlossenen Endsysteme, unabhängig von der Gruppenzugehörigkeit. Damit Endsystem 1 an alle Gruppenteilnehmer Daten senden kann, benötigt es die Adressen der beteiligten Endsysteme. Hierzu wird eine MARS_REQUEST-Nachricht an den MARS geschickt, der daraufhin auf dem ClusterControl VC mit der MARS_MULTI-Nachricht antwortet. Diese Nachricht enthält die angefragte Gruppenadresse und die Liste aller in der Gruppe angemeldeten Endsysteme. Anschließend kann Endsystem 1 eine Punkt-zu-Mehrpunkt-Verbindung zu den Teilnehmern aufbauen (SVC Setup) und Daten (Datentransfer) an die Gruppe senden. Das Abmelden geschieht mittels der MARS_LEAVE-Nachricht und verläuft analog zu einer Anmeldung.

Je nachdem, ob das VC-Mesh- oder das MCS-Schema angewendet wird, antwortet der MARS in der MARS_MULTI-Nachricht mit den Adressen aller Empfänger oder mit der Adresse des MCS. Es ist also für die Endsysteme kein Unterschied, welches Schema verwendet wird. Wie im obigen Beispiel gesehen, antwortet der MARS immer auf dem ClusterControl VC. Hierdurch erfahren alle angemeldeten Endsysteme sofort von Änderungen in den Gruppenzusammensetzungen und können somit ihre Verbindungen an die aktuelle Gruppensituation anpassen. Der Nachteil dieser Methode ist, dass viele Nachrichten von den Endsystemen verarbeitet werden müssen, obwohl sie für das Endsystem nicht relevant sind. Das ist z. B. der Fall, wenn das Endsystem nicht der bezeichneten Gruppe in der Nachricht angehört.

Signalisierung zwischen MARS und MCS

Beim MCS-Schema ist noch eine zusätzliche Signalisierung zwischen MARS und MCS notwendig. Der MCS muss über Gruppenänderungen vom MARS informiert werden.

Diese Signalisierung ist vom Ablauf her identisch mit der Signalisierung zwischen MARS und Endsystem. Statt eines ClusterControl VCs für die Endsysteme etabliert der MARS einen sogenannten ServerControl VC, um Gruppenänderungen an den MCS weiterzuleiten. Ein MCS kann sich beim MARS anmelden, und dabei angeben, für welche Gruppenadressen, bzw. für welchen Bereich im Gruppenadressraum er bereit ist, Gruppen zu bedienen. Es ist weiterhin auch vorgesehen, dass sich mehrere MCS beim MARS anmelden können. Was für Gruppenadressräume von den MCS bei der Anmeldung spezifiziert werden sollen, und wie der MARS mit mehreren MCS verfährt, ist allerdings nicht weiter in der MARS-Spezifikation festgelegt.

Meldet sich ein MCS für eine bereits aktive Gruppe an, die mittels dem VC-Mesh-Schema kommuniziert, leitet der MARS mit der MARS_MIGRATE-Nachricht einen Wechsel vom VC-Mesh- zum MCS-Schema ein. Die Nachricht wird auf dem ClusterControl VC an alle Endsysteme gesendet und bewirkt den Abbau aller diese Gruppe betreffenden Verbindungen bei den Endsystemen und anschließend den Aufbau einer neuen Verbindung zum MCS. Der umgekehrte Fall, die Abmeldung eines MCS, erfordert einen Wechsel vom MCS-Schema zum VC-Mesh-Schema. Das kann ohne Änderungen mit der MARS_LEAVE-Nachricht durchgeführt werden und ist transparent für die sendenden Endsysteme.

UNI 4.0

Das MARS-Konzept basiert auf der UNI3.1/3.0 Spezifikation [6]. Das ATM Forum hat in der aktuellen UNI Version 4.0 [21] die Unterstützung für Mehrpunkt-Verbindungen verbessert, indem sich Endsysteme bei einer Mehrpunkt-Verbindung als Empfänger anmelden können (Leaf Initiated Join). Hierzu benötigt das Endsystem die Ursprungsadresse und die Call-Reference-Nummer zur Identifizierung der Verbindung. Um diese Daten zu erhalten ist wiederum ein Server wie der MARS notwendig, der die Informationen über die Gruppenzugehörigkeiten der Teilnehmer registriert. Daher stellt die verbesserte Mehrpunkt-Unterstützung von UNI4.0 für die Emulation von IP-Multicast über ATM keine Verbesserung dar. Im Folgenden wird daher auch immer nur die UNI3.1/3.0 Spezifikation für Punkt-zu-Mehrpunkt-Verbindungen beachtet.

2.5. Zusammenfassung

Der Bereich Gruppenkommunikation umfasst eine Menge von zu beachtenden Eigenschaften und Aspekten. Ein Faktor, der bei fast allen Punkten eine Rolle spielt, ist die Skalierbarkeit. Hiermit wird bewertet, inwieweit die Qualität der Gruppenkommunikation mit zunehmender Gruppengröße und geografischer Ausdehnung abnimmt.

Die ATM-Technologie ist für Hochgeschwindigkeitsnetze entwickelt worden und bietet entsprechende Dienstqualitäten an. Das Konzept orientiert sich aber an einer verbindungsorientierten Unicast-Kommunikation und bietet nur eine rudimentäre Unterstützung für die Gruppenkommunikation in Form von Punkt-zu-Mehrpunkt-Verbindungen

an. Weitergehende Konzepte, wie Gruppenadressen und Multipeer-Kommunikation, finden hingegen keinerlei Berücksichtigung bei ATM.

Um dennoch eine Form der Gruppenkommunikation über ATM zu ermöglichen, ist eine Emulation von Mehrpunkt-zu-Mehrpunkt-Verbindungen notwendig. Das kann zum einen in der ATM-Schicht geschehen, wo eine effiziente Unterstützung für diesen neuen Verbindungstyp möglich wäre. Ein Nachteil dieser Vorgehensweise ist aber immer die notwendige Modifikation der bereits vorhandenen ATM-Infrastruktur in bestehenden Netzwerken. Zum anderen ist eine Emulation oberhalb der ATM-Schicht basierend auf den existierenden ATM-Verbindungsvarianten möglich. Hier bieten sich das VC-Mesh- und das MCS-Schema an. Beide Schemata haben allerdings Probleme bezüglich der Skalierbarkeit und sind nur sehr eingeschränkt für größere Gruppen geeignet. Für die Organisation dieser Schemata kann der MARS-Ansatz angewendet werden, der in diesem Bereich die bisher weiteste Verbreitung gefunden hat. Der MARS ist ein Konzept mit einem zentraler Server, der die Gruppen und deren Mitglieder verwaltet. Durch diese zentrale Organisationsform ist hier ebenfalls eine akzeptable Skalierbarkeit ausgeschlossen.

Ein weiterer Punkt, für den bisher keine adäquaten Lösungen gefunden worden sind, ist die Nutzung der Dienstgüteunterstützung von ATM für die Gruppenkommunikation. Das grundlegende Problem hierbei ist die Dynamik und der damit zusammenhängende schwankende Bandbreitenbedarf einer Gruppe und die Notwendigkeit bei ATM, die Dienstgüte im Vorhinein zu reservieren und nicht mehr ändern zu können.

3. Stand der Forschung

In diesem Kapitel werden aktuelle Forschungsarbeiten im Bereich Gruppenkommunikation über ATM vorgestellt. Das Ziel der meisten in diesem Kapitel vorgestellten Arbeiten ist es, eine Emulation von IP-Multicast über ATM zu ermöglichen. Hierbei sind zwei Bereiche zu unterscheiden, zum einen die Kommunikation innerhalb eines IP-Subnetzes (IGMP) und zum anderen Multicast-Routingprotokolle, die eine Subnetz-übergreifende Gruppenkommunikation ermöglichen. Einige wenige Arbeiten beschäftigen sich mit einer generischen Unterstützung der Gruppenkommunikation in ATM, unabhängig von IP-Multicast und Multicast-Routingprotokollen.

Anhand des ATM-Schichtenmodells in Abbildung 3.1 lassen sich die Forschungsarbeiten in zwei Kategorien unterteilen, in die Anwendungsschicht und die ATM-/AAL-Schicht. Die nebenstehende Tabelle 3.1 gibt eine Übersicht der dazugehörigen Forschungsarbeiten in den zwei Kategorien.

Die erste Kategorie sind die Forschungsarbeiten im MARS-Umfeld (siehe Kapitel 2.4). Bei MARS handelt es sich um ein Konzept zur Emulation von IP Multicast über ATM. Diese Forschungsarbeiten sind in der Anwendungsschicht angesiedelt, in der sich auch alle ATM-Anwendungen einordnen lassen. Diese Kategorie der Forschungsarbeiten behandelt im Wesentlichen eine grundsätzliche Emulation von IP-Multicast über ATM innerhalb eines IP-Subnetzes. Darüber hinaus werden auch einige Subnetz-übergreifende Lösungen vorgestellt und wie diese mit bestehenden Multicast-Routingprotokollen kooperieren.

Die zweite Kategorie sind Forschungsarbeiten, die eine mögliche Unterstützung der Gruppenkommunikation in der ATM- und AAL-Schicht behandeln. Hierzu zählt der Bereich der Signalisierung und das Multiplexen von verschiedenen ATM-Verbindungen. Um das Multiplexen zu vereinfachen, wird oftmals eine Unterstützung in der AAL-Schicht propagiert.

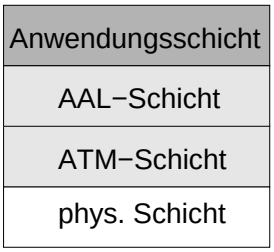


Abbildung 3.1.: Das ATM-Schichtenmodell.

Anwendungsschicht (MARS)	Kapitel (Seite)	ATM- und AAL-Schicht	Kapitel (Seite)
MARS[28]	2.4(23)	SMART[29]	3.3.1(46)
Bewertung des MARS[30]	3.2.1(32)	SEAM[31]	3.3.2(47)
VENUS[32]	3.2.2(34)	SPAM[33]	3.3.3(48)
EARTH[9]	3.2.3(36)	CRAM[34]	3.3.4(50)
Verteilter MARS[35]	3.2.4(37)		
MARS mit mehreren MCS[36]	3.2.5(39)		
PIM Sparse-Mode über ATM[10]	3.2.6(41)		
IP Multicast Shortcut Service[11]	3.2.7(42)		

Tabelle 3.1.: Einteilung der Forschungsarbeiten

Ein weiterer Bereich ist die Multicast-Unterstützung in den ATM-Schalteinheiten. Der Schwerpunkt dieses Gebietes sind Architekturen, die eine skalierbare Multicast-Unterstützung für das Switching von ATM-Zellen ermöglichen, wozu die Replikation und Weiterleitung von Zellen gehören. Aspekte der Gruppenkommunikation, wie z. B. Gruppenadressen oder Gruppenverwaltung werden darüber hinausgehend in diesem Bereich nicht betrachtet. Aus diesem Grund werden Forschungsarbeiten aus dem Bereich der ATM-Schalteinheiten in dieser Arbeit nicht behandelt.

Das Kapitel gliedert sich in drei Teile. Im Unterkapitel 3.1 werden die Kriterien vorgestellt, anhand derer die Forschungsarbeiten beurteilt werden. Hierauf folgt (Unterkapitel 3.2 und 3.3) die Vorstellung der einzelnen Arbeiten in den zwei Kategorien und am Kapitelende (Abschnitt 3.4) werden die vorgestellten Arbeiten zusammengefasst und anhand der vorgestellten Kriterien gegenübergestellt.

3.1. Beurteilungskriterien

Um die Arbeiten untereinander vergleichen zu können, sind eine Reihe von Kriterien ausgewählt, die die verschiedenen Aspekte und Anforderungen bzgl. der Gruppenkommunikation hervorheben. Die Kriterien lassen sich in zwei Kategorien aufteilen, die die wesentlichen Aspekte abdecken: Datentransport und Verwaltung. Die Kriterien sind im Einzelnen:

Datentransport: Diese Kriterien bewerten die Weiterleitung von Datenpaketen zwischen den Teilnehmern einer Gruppe.

- **Schema:** Gibt an, nach welchem Verteilschema die Benutzerdaten zwischen den Gruppenteilnehmern verteilt werden. In der Regel ist das Schema entweder VC Mesh oder MCS.
- **Verkehrskonzentration:** Wenn die an der Gruppenkommunikation aktiv beteiligten Netzwerkkomponenten durch Benutzerdaten ungleich belastet werden, kann dies zu Verzögerungen und Datenverlusten führen.

Datenformat: Können die Datenpakete vom Anwendungsprogramm direkt übernommen werden, oder müssen die Datenpakete zusätzlich aufbereitet werden (z. B. Einkapselung)?

Dienstgüteunterstützung: Beschreibt, ob bei dem Ansatz eine Dienstgüteunterstützung der ATM-Schicht berücksichtigt oder ermöglicht wird.

Verzögerung: Klassifiziert die Ende-zu-Ende-Verzögerung. Hier sind keine absoluten Werte angegeben, sondern nur qualitative Einschätzungen in Relation zu den anderen hier behandelten Arbeiten.

Robustheit: Dieses Kriterium bewertet, wie empfindlich das Verfahren gegenüber Ausfällen von Komponenten oder Übertragungsleitungen ist.

Ressourcenbedarf: Spezifiziert den logischen Ressourcenbedarf des Ansatzes. Das sind die benötigten ATM-Verbindungen. Dazu werden alle ankommenden und ausgehenden ATM-Verbindungen bei allen Gruppenteilnehmern gezählt. Eine 1:n-Verbindung hat demnach einen Ressourcenbedarf von $n + 1$ und eine 1:1-Verbindung einen Ressourcenbedarf von 2.

Verwaltung berücksichtigt die Komplexität der Gruppenverwaltung und den dadurch bedingten System-Overhead.

Organisation: Die grundsätzliche Organisationsform der Gruppenverwaltung. Hier wird hauptsächlich zwischen zentraler oder verteilter Verwaltung unterschieden. Bei der verteilten Verwaltung wird noch weiter differenziert, ob lokale Teilnehmer aggregiert und als ein Teilnehmer abstrahiert, oder ob alle lokalen Teilnehmerinformationen unverändert weitergegeben werden.

Ausfallsicherheit: Inwieweit wird die Verwaltung der Gruppenteilnehmer durch den Ausfall von Komponenten oder Verbindungsleitungen gestört und kann die Verwaltung diese Ausfälle kompensieren?

Signalisierungsaufwand: Beziffert die durchschnittliche Anzahl der Komponenten, die bei einem Teilnehmerbeitritt oder -austritt an der Signalisierung beteiligt sind.

IDMR-Protokolle: **I**nter-**D**omain-**M**ulticast-**R**outing-Protokolle. Gibt an, ob der Ansatz die Kooperation mit bestehenden IP-Multicast-Routingprotokollen berücksichtigt.

3.2. Anwendungsschicht (MARS)

Grundsätzlich bietet ATM für Anwendungen bidirektionale 1-1 (Unicast) Verbindungen und unidirektionale 1-n (Multicast) Verbindungen an. Eine Unterstützung für m-n (Multipeer) oder m-1 (Concast) Verbindungen ist jedoch nicht vorhanden (siehe Grundlagen, Unterkapitel 2.2, ab Seite 11). Die Aufgabe der Anwendungsschicht besteht im Wesentlichen darin, mittels den vorhandenen ATM-Verbindungstypen (1:1 oder 1:n)

Gruppenkommunikation in Form einer Multipeer-Verbindung zu emulieren. Hierzu gibt es grundsätzlich die beiden Möglichkeiten VC Mesh und MCS, die schon beim MARS-Konzept im Grundlagenkapitel 2.4 vorgestellt worden sind.

3.2.1. Bewertung des MARS

Die Reichweite eines MARS Dienstes wird als MARS Cluster bezeichnet. Das ist das ATM-Netz, oder ein Teil des ATM-Netzes, für den ein MARS zuständig ist. In der Regel ist dabei der MARS Cluster mit einem IPv4 Logical IP Subnet (LIS, siehe auch Unterkapitel 2.3) identisch, da ein LIS meist auch die Administrationsgrenzen widerspiegelt. Die Größe eines MARS Clusters unterliegt darüber hinaus noch einer Reihe von Bedingungen, die im Folgenden vorgestellt werden [30].

Ein MARS Cluster ist aus Netzwerksicht eine Menge von ATM-Schnittstellen, die IP-Multicast Daten direkt über ATM-Verbindungen weiterleiten. Jede Schnittstelle muss dazu Zustandsinformationen über die zugehörigen Gruppen und Empfänger speichern und regelmäßig diesen Zustand aktualisieren.

Die Größe eines Clusters hat dabei zwei Bedeutungen: Zum einen die Anzahl der Endsysteme, die einen MARS nutzen können und zum anderen die geografische Verteilung der Endsysteme. Die Anzahl der Endsysteme hat Auswirkungen auf die Menge der Zustandsinformationen und auf die mittlere Rate der Signalisierungsnachrichten, die vom MARS verteilt werden. Mit der Anzahl der Signalisierungsnachrichten steigt auch die Anzahl der zu modifizierenden ATM-Verbindungen. Der zweite, die Größe beeinflussende Faktor ist die geografische Verteilung der Endsysteme. Dieser Faktor hat Auswirkungen auf die Verzögerung der Signalisierungsnachrichten und auf die Dauer des Auf- und Abbaus von ATM-Verbindungen.

Um die Begrenzungen bei der Größe eines MARS-Clusters zu bestimmen, werden im Folgenden Szenarien angenommen, die den ungünstigsten Fall darstellen. Der Großteil der Analysen setzt dazu VC Mesh basierte Gruppen voraus, wobei alle Mitglieder im Cluster gleichzeitig Sender und Empfänger sind. Der Einsatz eines MCS verringert die Begrenzungen, und ermöglicht größere Cluster. Die Begrenzungen ergeben sich aus den folgenden Randbedingungen:

1. Die maximale Anzahl von ATM-Verbindungen, die ATM-Schnittstellenkarten erzeugen oder annehmen können ($VCmax$).

2. Die maximale Anzahl von Blattknoten die ein Wurzelknoten bei einer Punkt-zu-Mehrpunkt-Verbindung verwalten kann ($LEAFmax$). Die UNI-Signalisierung erlaubt maximal $2^{15} = 32768$ Endpunkte, also $LEAFmax <= 32768$.

3. Die durch Gruppenänderungen bedingten Belastungen im Netzwerk und den Endsystemen.

4. Durch die geografische Ausdehnung des Clusters bedingte Verzögerungen.

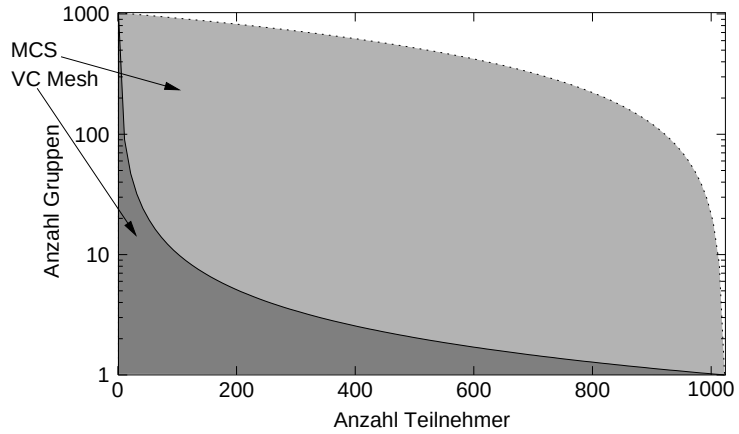


Abbildung 3.2.: Größenbegrenzung im MARS Modell in Abhängigkeit von Gruppen- und Teilnehmergröße bei VC Mesh und MCS.

Des Weiteren wird im Folgenden durchgehend von UBR (Best Effort) Verbindungen ausgegangen, die keinerlei Bandbreiten auf den Leitungen oder in einer ATM-Schalteinheit reservieren. Bei anderen Dienstgütekategorien kämen noch weitere begrenzende Faktoren hinzu.

Ein zu UBR ähnlicher Diensttyp ist ABR. Hiermit könnten die im Netzwerk vorhandenen Ressourcen besser auf die ATM-Verbindungen aufgeteilt und Stausituationen vermieden werden. Trotz der Zweckmäßigkeit von ABR gibt es Probleme beim Einsatz von ABR: Die Unterstützung von ABR für Punkt-zu-Mehrpunkt-Verbindungen in den ATM-Schalteinheiten ist nicht immer vorhanden und viele Endsystemen haben ATM Netzwerkkarten (und Treiber) die ABR nicht unterstützen.

Sei N die Anzahl der Endsysteme im MARS Cluster und G die Anzahl der aktiven Gruppen, dann gibt es für VC Mesh und MCS folgende Begrenzungen:

VC Mesh	MCS
$N \leq LEAFmax$	
$NG + 2 \leq VCmax$	$N + G + 2 \leq VCmax$

Bewertung

Abbildung 3.2 zeigt beispielhaft für $VCmax = LEAFmax = 1024$ den Zusammenhang zwischen Teilnehmeranzahl und Gruppengröße für VC Mesh und MCS. In Abhängigkeit von der Anzahl der Teilnehmer ist die Anzahl der möglichen Gruppen für das VC Mesh und das MCS-Schema eingezeichnet. Die Anzahl der Gruppen ist dabei logarithmisch dargestellt, um eine bessere Differenzierung zu ermöglichen. Wie deutlich zu erkennen ist, kann mit dem MCS-Schema eine höhere Clustergröße ermöglicht werden. Hieraus ergibt sich, dass das MCS Schema einen geringeren **Ressourcenbedarf** hat, als das VC-Mesh-Schema. Durch die **Verkehrskonzentration** im MCS kann die Anzahl der benötigten ATM-Verbindungen signifikant reduziert werden, wodurch aber auch die **Verzögerung** im MCS im gleichen Maße zunimmt.

Die potenzielle Belastung durch den **Signalisierungsaufwand** steigt ebenfalls mit der Anzahl N der Endsysteme. Die mittlere Frequenz von Join/Leave-Signalisierungsnachrichten, die der MARS zu verarbeiten und weiterzuleiten hat, nimmt mit der Anzahl der Gruppenteilnehmer zu. Diese Zunahme hat auch Auswirkungen auf die Signalisierungsaktivität in der ATM-Schicht, da die Sender aufgrund von Join/Leave-Signalisierungsnachrichten beteiligte Empfänger mittels ADD_PARTY- und DROP_PARTY-Signalisierungsnachrichten zu einer ATM-Verbindung hinzufügen oder entfernen müssen.

3.2.2. VENUS

Das MARS-Modell ist prinzipiell unabhängig von einem LIS. Aus praktischen Erwägungen wird ein MARS aber in der Regel immer nur innerhalb eines LIS eingesetzt (intra-LIS). Zwischen benachbarten LIS (bzw. MARS Clustern) werden IP-Multicast Pakete mittels eines Multicast-Routers weitergeleitet (inter-LIS). Das geschieht in einer ähnlichen Weise zu 'Classical IP over ATM', wo ein IP-Router Unicast-Pakete zwischen den LIS weiterleitet. Die Entwicklung von NHRP [37] (**N**ext **H**op **R**esolution **P**rotocol) als Unicast-Shortcut Mechanismus zur Überbrückung von LIS-Grenzen in ATM-Netzen hat die Frage nach einem ähnlichen Mechanismus für IP-Multicast in ATM-Netzen aufgeworfen.

VENUS steht für **V**ery **E**xtensive **N**on-**U**nicast **S**ervice [32] und ist von G. Armitage, dem Autor des MARS-Modells, verfasst worden. VENUS ist eine hypothetische Lösung, anhand derer die Probleme einer Ausweitung des MARS-Modells über mehrere LIS diskutiert werden. Das Dokument zeigt, welch hohe funktionale Komplexität so ein Dienst haben müsste, und stellt die Anforderungen an einen solchen Dienst heraus. Eine konkrete Lösung für die Problematik wird hingegen nicht behandelt.

Eine VENUS-Domäne ist definiert als eine Menge von zwei oder mehr beteiligten logischen MARS Clustern. Es wird bei VENUS aber kein Protokoll angegeben, das die Kommunikation zwischen den MARS Clustern beschreibt. Eine Multicast-Shortcut-Verbindung ist dann eine Punkt-zu-Mehrpunkt-Verbindung, deren Empfängerknoten in der VENUS-Domäne verteilt sind. Diese Shortcut-Verbindung ist dabei nicht an LIS-Grenzen gebunden und kann somit evtl. vorhandene Multicast-Router umgehen. VENUS behandelt dabei das Erste von zwei grundlegenden Problemen:

- Das erste Problem ist der stark erhöhte Umfang der Endsysteme, den individuelle ATM Schnittstellen beobachten und dabei auf Gruppenmitgliedschaftsänderungen reagieren müssen. Bei IP-Multicast übernehmen Multicast-Router diese Aufgabe als Aggregationspunkte für Datenpakete und Signalisierungsnachrichten über Gruppenänderungen. Innerhalb einer VENUS-Domäne mit Shortcut-Verbindungen existieren diese Aggregationspunkte nicht mehr. Daraus folgt, dass alle Quellen innerhalb einer Domäne alle Gruppenänderungen beachten müssen. Hieraus ergibt sich, dass eine VENUS-Domäne einem großen MARS Cluster bezüglich der Begrenzungen der Größe sehr ähnlich ist.

- Das zweite Problem sind die Auswirkungen von Shortcut-Mechanismen auf Inter Domain Multicast Routing (IDMR) Protokolle, wie z. B. DVMRP (Distance Vector

Multicast Routing Protocol, [38]) . Eine Gruppe hat eine Vielzahl von Sendern und Empfängern, verteilt über die einzelnen Cluster. IDMR-Protokolle berechnen einen effizienten Inter-Domain Multicast-Baum zwischen den beteiligten Multicast-Routern, indem sie Gruppenteilnehmer zusammenfassen. Wenn aber Quellen einer Gruppe einfach durch Shortcut-Verbindungen Multicast-Router umgehen können, ist es notwendig, dass jede dieser Shortcut-Verbindungen dem IDMR-Protokoll mitgeteilt wird, so dass das Protokoll diese Verbindungen in die Berechnung mit einbeziehen kann. Des Weiteren kann eine Anpassung des IDMR-Protokolls nötig sein (siehe hierzu Unterkapitel 3.2.6 und 3.2.7).

VENUS ist ein hypothetischer Mechanismus um die beteiligten MARS Cluster zu koordinieren. Bei VENUS werden nur die Randbedingungen diskutiert, die ein Shortcut-Mechanismus zu erfüllen hat. Alle darüber hinausgehenden Anforderungen werden von VENUS nicht beachtet. Es wird zur Vereinfachung vorausgesetzt, dass jeder MARS über alle Daten zur Unterstützung von Teilnehmeränderungen informiert wird, so dass die ATM-Verbindungen entsprechend aufgebaut und verwaltet werden können. Innerhalb einer VENUS-Domäne wird jedes Endsystem über die Gruppenmitglieder (und deren Änderung) aller teilnehmenden MARS Cluster informiert. Die hieraus entstehende Problematik teilt sich in zwei Bereiche auf: den Verbindungsaufbau und das Verbindungsmanagement:

Verbindungsaufbau: Möchte ein neuer Sender an eine Gruppe senden, so müssen alle beteiligten MARS Cluster über das Vorhandensein dieser Gruppe befragt werden. Eine mögliche Optimierung wäre, dass diese Anfrage nur entlang des IP-Multicast-Baums weitergeleitet wird. Das Ergebnis ist in jedem Fall, dass der neue Sender über alle Gruppenmitglieder informiert wird. Um das zu verhindern, könnte die Reichweite der Anfrage begrenzt werden. Das hätte dann zur Folge, dass Gruppenmitglieder in der Nähe direkt antworten und entfernte Mitglieder über einen Multicast-Router zusammengefasst werden. Dieser Mechanismus würde aber nur in einfachen Netzwerktopologien funktionieren, z. B. eine lineare Aneinanderreihung von Clustern. In komplexeren Topologien würde die Anfrage sehr oft verzweigen und es würden immer noch eine Vielzahl von Teilnehmern antworten.

Ein anderes Problem ist, dass die verschiedenen Cluster entweder das VC-Mesh- oder das MCS-Schema verwenden können. Beim MCS-Schema muss VENUS dafür sorgen, dass nur die MCS-Adresse weitergegeben wird und nicht die Adressen aller Teilnehmer im MCS-basierten Cluster.

Verbindungsmanagement: Ist einmal eine ATM-Verbindung von einem Sender zu den beteiligten Empfängern der Gruppe aufgebaut, muss diese Verbindung immer auf dem neuesten Stand gehalten werden. Die Konsequenz hieraus ist, dass jede Empfängeränderung an alle Sender der Gruppe weitergeleitet werden muss. Die daraus resultierende Signalisierungsbelastung ist nicht lokal auf einen Cluster begrenzt, sondern erstreckt sich auf die gesamte VENUS-Domäne.

Bewertung

VENUS hat die gleichen Beschränkungen bzgl. der Größe wie MARS (siehe Unterkapitel 3.2.1) und somit auch den gleichen **Ressourcenbedarf**. Die Anzahl der Teilnehmer wird durch die maximal mögliche Zahl von ATM-Verbindungen begrenzt. Des Weiteren entsteht ein hoher **Signalisierungsaufwand** durch die Verwaltung. Insbesondere ist im Endsystem bei der Reassemblierung der ATM-Zellen zu Paketen von einer erhöhten **Verzögerung** und Endsystembelastung auszugehen, wenn eine große Zahl ankommender Verbindungen existiert.

3.2.3. EARTH

EARTH steht für **EA**sy IP multicast **R**outing **TH**rough ATM clouds [9] und versteht sich konzeptionell zwischen MARS (Kapitel 2.4) und VENUS (Kapitel 3.2.2) angesiedelt.

EARTH führt, analog zur VENUS-Domäne, das Konzept des Multicast LIS (MLIS) ein. Dieses Multicast-Subnetz ist LIS-übergreifend und kann das gesamte physikalische ATM-Netz umfassen. Innerhalb des MLIS werden IP-Multicast-Adressen aufgelöst und zwischen den Rechnern im MLIS ist somit ein Shortcut-Routing möglich. Im Gegensatz zum VENUS-Konzept steht allerdings nur ein zentraler EARTH-Server (äquivalent zum MARS) für die Adressauflösung im MLIS zur Verfügung. Der Mechanismus zur Adressauflösung unterscheidet sich allerdings vom MARS Konzept. Die Adressen der Gruppenmitglieder werden nicht wie bei MARS explizit vom EARTH-Server propagiert. Der EARTH-Server speichert nur die teilnehmenden Empfänger einer Gruppe. Die Sender müssen periodisch die Empfängerliste vom EARTH-Server abfragen und die notwendigen Veränderungen selbstständig ermitteln.

Innerhalb eines MLIS wird das VC-Mesh-Schema verwendet, um Daten zwischen den Gruppenmitgliedern auszutauschen. Zwischen benachbarten MLIS und als Ausgangsrouter am Rand des ATM-Netzes kommen IP-Multicast-Router zum Einsatz. Hierzu gehört zu jedem EARTH-Server ein sogenannter e-mrouted (EARTH Multicast Routing Daemon). Dieser verbindet die einzelnen Multicast-Router an der IP/ATM-Grenze und sorgt für eine direkte ATM-Verbindung zwischen diesen EARTH-Multicast-Routern. Sind im MLIS ebenfalls aktive Teilnehmer für Gruppen vorhanden, die auch bei den Multicast-Routern aktiv sind, so werden entsprechende IGMP-Nachrichten [15] an den jeweiligen EARTH-Multicast-Router gesendet.

Im EARTH-Konzept ist eine rudimentäre Dienstgüteunterstützung auf der Basis von Dienstgüteebenen vorgesehen. Empfänger können sich für im Vorhinein festgelegte Dienstgüteebene anmelden. Diese müssen allerdings auch von den Sendern explizit unterstützt werden. Für jede Dienstgüteebene werden separate ATM-Verbindungen von den Sendern zu allen Empfängern aufgebaut. Dieses Konzept wird realisiert, indem für die unterschiedlichen Dienstgüteebenen verschiedene Empfängerlisten im EARTH-Server vorgehalten werden.

Um für die Ausfallsicherheit eines EARTH-Servers zu sorgen, soll das Server Cache Synchronisation Protocol (SCSP) [39] (siehe auch Unterkapitel 3.2.4) eingesetzt werden. Alle im ATM-Netz vorhandenen EARTH-Server (die einer SCSP-Server-Gruppe ange-

hören) koordinieren ihre Zustände untereinander. Bei Ausfall eines EARTH-Servers wird dieser durch den Server eines benachbarten MLIS ersetzt.

Bewertung

Der EARTH-Ansatz stellt eine mögliche Realisierung des VENUS-Konzeptes dar, der zusätzlich noch die Problematik der Interaktion zwischen Shortcut-Verbindungen und Multicast-Routingprotokollen (**IDMR-Protokolle**) behandelt. Zusätzlich wird eine rudimentäre **Dienstgüteunterstützung** im ATM-Netz angeboten und durch den Einsatz von SCSP kann die **Ausfallsicherheit** durch Backup-Systeme erhöht werden. Nichtsdestotrotz verwendet EARTH dieselben Schemata wie MARS, um die Daten innerhalb einer Gruppe zu verteilen und stößt somit auf die gleichen Grenzen bzgl. Gruppengröße (**Ressourcenbedarf**), Skalierbarkeit und **Verzögerung** wie MARS und VENUS. Das EARTH-Konzept behandelt eher Probleme der Gruppenverwaltung und enthält keine Verbesserungen für den Datentransport.

3.2.4. Verteilter MARS

Der MARS-Server ist eine zentrale Komponente für die Gruppenkommunikation innerhalb eines LIS. Um eine robuste (unempfindlich gegen Ausfälle von Komponenten) Gruppenkommunikation zu ermöglichen, ist es daher notwendig, Sicherungssysteme für den MARS-Server vorzusehen.

Eine Grundvoraussetzung für einen verteilten MARS [35] ist die Synchronisation und Replikation der Datenbestände der MARS-Server. Für eine Synchronisation von verteilten Datenbeständen ist von der IETF Arbeitsgruppe 'Internetworking over NBMA' (Non-Broadcast-Medium-Access) das **S**erver **C**ache **S**ynchronisation **P**rotocol (SCSP) [39] entwickelt worden. Hiermit können z. B. Next Hop Server (Komponente von NHRP, [37]) oder MARS-Server synchronisiert werden. Aus diesem Grund wird zunächst das Protokoll SCSP näher erläutert und anschließend der Einsatz von SCSP beim Konzept für einen verteilten MARS vorgestellt.

SCSP

SCSP stellt einen generalisierten Mechanismus zur Verfügung um Probleme bei Server-Cache-Synchronisation und -Replikation zu lösen. SCSP synchronisiert Caches (oder Teile von Caches) zwischen einer Menge von Servern. Diese Server werden in einer sogenannten Server Group (SG) zusammengefasst, innerhalb derer alle Server Ihre Datenbestände untereinander abgleichen.

Bei SCSP werden die gleichen Mechanismen zur Synchronisation angewandt wie bei OSPF [40] (Hello-, Datenbanksynchronisations- und Flooding-Prozeduren). SCSP besteht aus drei Phasen: Die erste Phase ist die Hello-Phase, in der zwei Teilnehmer feststellen, dass sie in der gleichen Server Group sind. Darauf folgt in der zweiten Phase die Synchronisation der Datenbestände, so dass benachbarte SCSP-Server dieselben

Datenbestände haben. In der dritten Phase (flooding state) werden nur noch Aktualisierungen der Datenbestände ausgetauscht, so dass jede Änderung innerhalb einer Server Group weitergegeben wird. Abbildung 3.3 zeigt das Zustandsübergangsdiagramm von SCSP und die darin angesiedelten drei Phasen des Protokolls.

Als eigenständiges Protokoll ist SCSP nicht verwendbar, da es keinerlei Annahmen über die Form und Eigenschaften der Datenbestände macht. Es muss mit einem Protokoll verbunden werden, das die Mechanismen von SCSP einsetzt und die von SCSP gelieferten Datenbestände interpretiert. Dieses Protokoll muss dann auch spezifizieren, welche Datenbestände zu welchen Zeitpunkten zwischen den Servern ausgetauscht und aktualisiert werden. SCSP ist nur für den eigentlichen Datenaustausch zuständig.

Ausfall eines MARS-Servers

Der Einsatz von SCSP zwischen mehreren MARS-Servern soll im Wesentlichen einen fehlertoleranten Betrieb ermöglichen. Um das zu erreichen, gibt es innerhalb eines MARS Clusters (i. Allg. identisch mit einem LIS) mehrere MARS-Server, wovon aber immer nur einer aktiv ist. Die MARS-Server innerhalb des Clusters bilden dann eine SCSP Server Group (SG). Die weiteren Server dienen nur als Backup und werden erst nach Ausfall des primären MARS-Servers aktiv.

Bei Ausfall des aktiven MARS-Servers ist es Aufgabe der Endsysteme sich an einen neuen MARS-Server anzubinden. Hierzu muss bei allen Endsystemen mindestens ein entsprechender Backup-MARS-Server eingetragen sein. Dieser eingetragene Backup-Server muss dabei auch bei allen Endsystemen identisch sein. Die Endsysteme bauen eine Signalisierungsverbindung zum neuen MARS-Server auf (Abbildung 3.4) und der MARS-Server etabliert eine Signalisierungsverbindung zu den Endsystemen und dem MCS (CCVC und SCVC, siehe Kapitel 2.4). Daraufhin kann der neue MARS-Server die Verwaltung der IP-Multicast-Gruppen übernehmen und die Aufgaben des ausgefallenen MARS-Servers fortsetzen.

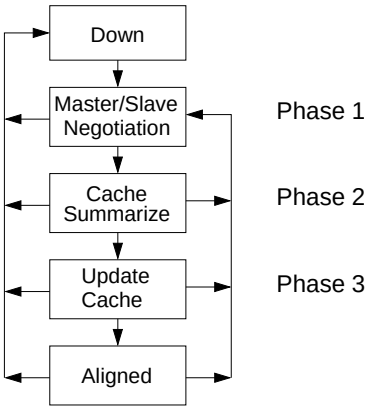


Abbildung 3.3.: Zustandsübergangsdiagramm von SCSP.

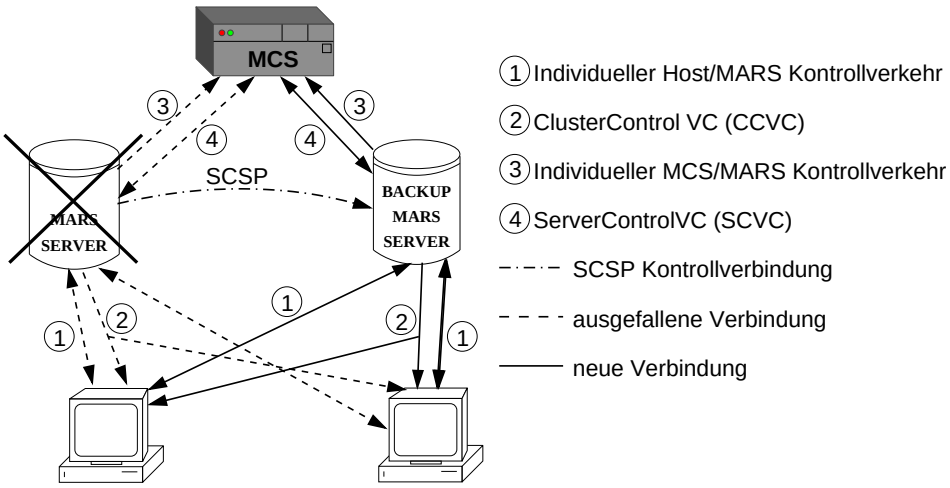


Abbildung 3.4.: Ausfall eines MARS-Servers und Migration zum Backup-Server.

Bewertung

Dieser Ansatz behandelt im Wesentlichen nur den Bereich der **Ausfallsicherheit**. Um diese zu erhöhen werden Backup-Systeme eingesetzt, die mit dem Protokoll SCSP synchronisiert werden. Aspekte der **Fehlertoleranz** beim **Datentransport** werden nicht erläutert. Die Backup-Systeme müssen als zusätzliche Komponenten realisiert sein und benötigen somit zusätzliche Ressourcen. Darüber hinaus bleibt die zentrale **Organisation** der Verwaltung des MARS erhalten und dadurch auch die Nachteile dieser Organisationsform.

3.2.5. MARS mit mehreren MCS

Ein kritisierter Nachteil des MARS-Modells ist der mögliche Engpass im MCS bei vielen aktiven Sendern. Eine Gruppe ist immer an einen MCS gebunden, und sobald die Anzahl der Sender in der Gruppe wächst, wird der MCS zum Flaschenhals und die Ende-zu-Ende-Paketverzögerung erhöht sich mit zunehmender Senderanzahl [41]. Im RFC 2149 [36] wird ein Mechanismus beschrieben, wie mehrere MCS pro Multicast-Gruppe für einen fehlertoleranten Betrieb eingesetzt werden können. Eine Erweiterung des Ansatzes, um eine Lastverteilung auf vorhandene MCS durchzuführen wird in [42] beschrieben.

Fehlertoleranter Betrieb mit mehreren MCS

Im MARS Cluster können eine Menge von MCS vorhanden sein. Diese MCS sind dabei in einer Liste sortiert, welche allen beteiligten MCS bekannt sein muss. Der erste MCS in dieser Liste ist der aktive MCS, der alle Gruppen, wie beim MARS-Konzept beschrieben, bearbeitet. Der aktive MCS hat zusätzlich eine Punkt-zu-Mehrpunkt-Verbindung zu den anderen MCS in der Liste geöffnet (den sogenannten HelloVC). Über diese Verbindung werden periodisch KeepAlive-Nachrichten verschickt. Einen Ausfall des aktiven MCS wird den anderen MCS indirekt durch das Ausbleiben der KeepAlive-Nachrichten

signalisiert. Ist das der Fall, wird der nächste MCS in der Liste aktiv. Dieser meldet sich beim MARS an und baut ebenfalls einen HelloVC zu den verbliebenen MCS auf. Die Anmeldung beim MARS bewirkt, dass dieser alle MARS Clients über den neuen MCS informiert.

Ein einmal ausgefallener MCS kann in diesem Konzept nicht wieder aktiviert werden. Des Weiteren haben die verbliebenen MCS keine aktive Funktion in diesem Konzept. Eine Lastverteilung zwischen den vorhandenen MCS findet nicht statt. Das Konzept hat den Vorteil, dass nur Änderungen auf Seiten des MCS notwendig sind, jedoch keine Änderungen beim MARS-Server und den Endsystemen.

Lastverteilung zwischen mehreren MCS

Eine konzeptionelle Erweiterung dieses Ansatzes um eine Lastverteilung wird in [42] vorgestellt. Eine Voraussetzung für die Lastverteilung ist, dass mehrere MCS aktiv eine Gruppe unterstützen, indem die Sender oder Gruppenmitglieder unter den MCS aufgeteilt werden. Einer der aktiven MCS, der sogenannte primäre MCS übernimmt dabei die zusätzliche Aufgabe als Synchronisationseinheit der anderen (sekundären) MCS. Dafür ist ein Protokoll spezifiziert, mit dem die sekundären MCS dem primären MCS signalisieren, welche Gruppen, bzw. Sender von Gruppen sie bedienen wollen.

Eine geregelte Lastverteilung ist bei diesem Vorgehen nicht vorhanden, es wird nur beschrieben, wie die Last auf mehrere MCS aufgeteilt werden kann, ohne dass Änderungen beim MARS-Server nötig sind.

Lastverteilung mit dem MARS-Server

Mit einer Erweiterung des MARS-Servers (ebenfalls [42]) ist hingegen eine erheblich bessere Lastverteilung möglich. Zum einen sind keine Änderungen am MCS nötig (die oben beschriebene Kommunikation zwischen den MCS entfällt), und zum anderen hat der MARS-Server alle Informationen über die Gruppen und deren Mitglieder. Mit diesen Informationen ist der MARS-Server wesentlich besser geeignet, die Gruppen und Sender auf die MCS zu verteilen. Darüber hinaus wird zum Punkt Lastverteilung nur angegeben, dass die Sender und Gruppen vom MARS-Server gleichmäßig auf die vorhandenen MCS verteilt werden sollen. Eine Berücksichtigung des tatsächlichen Datenverkehrs findet nicht statt. Diese Art der Lastverteilung kann nur mit der Annahme gerechtfertigt werden, dass alle Gruppen im Mittel in etwa gleich groß sind und ein ähnliches Datenaufkommen haben. In Anbetracht verschiedener Anwendungstypen (z. B. Video, Audio, Whiteboard) scheint diese Annahme nicht gerechtfertigt zu sein.

Bewertung

Die hier vorgestellten drei Möglichkeiten, das MARS-Konzept mit mehreren MCS zu erweitern, erfüllen immer nur eingeschränkt die Anforderungen. Die demgegenüber propagierten Vorteile (keine Änderungen beim MARS oder beim MCS), können die gemachten Einschränkungen nicht kompensieren. Bezogen auf den **Datentransport** ist der Einsatz mehrerer MCS von Vorteil, da hierdurch die **Verkehrskonzentration** im MCS

gemindert werden kann, was auch Einfluss auf eine geringere **Verzögerung** hätte. Das ist konkret bei den Ansätzen nicht zu garantieren, da das tatsächliche Datenaufkommen bei der Lastverteilung nicht beachtet wird.

Die Ansätze steigern die **Fehlertoleranz**, erfordern aber auch einen höheren Ressourcenbedarf, da mehr ATM-Verbindungen benötigt werden. Aus Sicht der **Verwaltung** wird die **Ausfallsicherheit** ebenfalls nicht erhöht, da nur ein MARS existiert. Der **Signalisierungsaufwand** bei diesen Ansätzen ist höher, da mehrere MCS involviert sind.

3.2.6. Unterstützung für PIM Sparse-Mode über ATM

Im RFC 2337 [10] wird eine mögliche Unterstützung des Multicast-Routingprotokolls PIM Sparse-Mode [13] im ATM-Backbone vorgestellt. Die Unterstützung ist dabei vollkommen unabhängig vom MARS-Modell oder anderen IP-Multicast Emulationen. Das Multicast-Routingprotokoll PIM Sparse-Mode basiert auf der Annahme, dass die Systeme eher weit voneinander entfernt sind und somit nur eine geringe Dichte von Gruppenteilnehmern vorhanden ist (im Gegensatz zu PIM Dense-Mode [14]). Ein fester Bestandteil von PIM Sparse-Mode ist der explizite Gruppenbeitritt und die Bereitstellung von Rendezvous-Stellen. An eine Rendezvous-Stelle werden alle Daten einer Gruppe gesendet, und dann von dort an alle Empfänger verteilt.

Für die Unterstützung von PIM Sparse-Mode im ATM-Netz wird vorausgesetzt, dass alle am ATM-Netz angeschlossenen PIM-Router einem LIS angehören und dass alle PIM-Router die IP- und ATM-Adressen der anderen beteiligten PIM-Router kennen. Ob das mit Hilfe eines ATMARP-Servers [23] oder über eine statische Konfiguration erfolgt, wird nicht weiter festgelegt. Die Kommunikation über mehrere LIS und die Anbindung von Endsystemen innerhalb eines LIS an PIM Sparse-Mode wird ebenfalls nicht behandelt.

Jeder PIM-Router hat eine ATM-Punkt-zu-Mehrpunkt-Verbindung zu allen anderen PIM-Routern (entspricht dem VC-Mesh-Schema). Über diese Verbindung sendet der Router sämtliche Multicast-Pakete, sowohl Daten- als auch Kontrollverkehr. Zusätzlich können die PIM-Router für Gruppen individuelle ATM-Verbindungen aufbauen, wodurch unnötiger Datenverkehr zu nicht beteiligten PIM-Routern vermieden werden kann. Hierzu entscheidet jeder PIM-Router lokal, ob die Daten einer Gruppe über eine gruppenspezifische Verbindung geschickt werden. Kriterien, die für die Etablierung einer gruppenspezifischen Verbindung berücksichtigt werden sollten, sind die Verkehrsmenge der Gruppe und der Verzweigungsfaktor, also die angemeldeten PIM-Router der Gruppe.

Entscheidet sich einer der PIM-Router im LIS für einen Sender einen spezifischen Baum zu etablieren, wie es in der PIM Sparse-Mode Spezifikation vorgesehen ist, so baut dieser Router als Wurzel eine Punkt-zu-Mehrpunkt-Verbindung für diesen Sender auf. Alle anderen beteiligten PIM-Router im LIS müssen dann ebenfalls zu dem spezifischen Baum für den Sender wechseln. Da ein PIM-Router, der den Wechsel veranlasst, PIM Join(S,G) Nachrichten sendet, sind alle anderen PIM-Router im LIS in der Lage, diese Information zu empfangen. Beim Hinzufügen eines neuen Teilnehmers (PIM-Router im LIS) zu einem senderspezifischen Baum muss die Wurzel über diesen neuen Teilnehmer informiert werden. Hierzu sendet der PIM-Router zuerst (*,G)-Join

Nachrichten in Richtung der Rendezvous-Stelle auf der gemeinsamen ATM-Verbindung. Diese Nachrichten werden von der Wurzel empfangen, und der neue PIM-Router wird dem senderspezifischen Baum hinzugefügt.

Um das Problem von Paketreflektionen (siehe hierzu Unterkapitel 2.4, ab Seite 23) zu vermeiden, darf ein PIM-Router nur Pakete außerhalb des LIS an andere Router im LIS weiterleiten. Mit diesem einfachen Mechanismus können Reflektionen verhindert werden, und es ist nicht nötig, wie beim MARS, die Pakete extra einzukapseln.

Bewertung

Im Gegensatz zum MARS ist dieser Entwurf in der **Verwaltung** sehr einfach gehalten und er besitzt auch nicht die Flexibilität vom MARS-Protokoll. Für eine konkrete Umsetzung sind hier viele Bestandteile noch nicht spezifiziert, wie die Kenntnis der IP- und ATM-Adressen. Andererseits wird auch kein zentraler Server für die **Organisation** der Verwaltung benötigt. Der **Signalisierungsaufwand** ist moderat und orientiert sich am PIM Sparse-Mode Protokoll, es werden keine zusätzlichen Nachrichten benötigt.

Für den **Datentransport** wird das VC-Mesh-**Schema** eingesetzt, allerdings unabhängig von Gruppen. Das kann zu einem erhöhten Datenaufkommen führen, es ist aber die Möglichkeit gruppenspezifischer Verbindungen vorgesehen. Dabei wird aber nicht konkret angegeben, wann diese Verbindungsart gewählt wird. Um eine intra-LIS Gruppenkommunikation über ein ATM-Backbone zwischen PIM-Routern zu etablieren, sind die Mechanismen dieses Ansatzes aber ausreichend.

3.2.7. IP Multicast Shortcut Service (IMSS)

IMSS [11] ist ein Ansatz, der für große ATM-Netze (ATM WANs) ein Shortcut-Routing für IP Multicast anbietet. Das Shortcut-Routing wird für Multicast-Router zwischen verschiedenen LIS (inter-LIS) etabliert. Innerhalb eines LIS (intra-LIS) wird weiterhin auf den Einsatz von MARS verwiesen. IMSS basiert auf zwei verschiedenen Komponenten: CONGRESS (**CON**nection-oriented **G**roup address **RES**olution **S**ervice) ist für die Auflösung einer IP-Multicast-Adresse in eine Menge von Multicast-Router-Adressen verantwortlich, die den an diese Multicast-Adresse gerichteten Datenverkehr empfangen sollen. Die zweite Komponente, IP-SENATE (**IP** multicast **SE**rvice for **N**on-broadcast **A**ccess **N**etworking **TE**chnology), ist für die Routingentscheidungen und die Datenübertragung zuständig. Die beiden Komponenten werden im Folgenden getrennt erläutert.

CONGRESS

CONGRESS bietet ein generisches Protokoll für die dynamische Adressauflösung von IP-Multicast-Adressen in eine Menge von Multicast-Router-Adressen, die an ein ATM-Netzwerk angeschlossen sind. Die aufgelösten Adressen werden dann im nächsten Schritt von IP-SENATE verwendet, um eine Shortcut-Kommunikation zwischen den Multicast-Routern zu etablieren. CONGRESS kann allgemein für NBMA-Netze (Non Broadcast

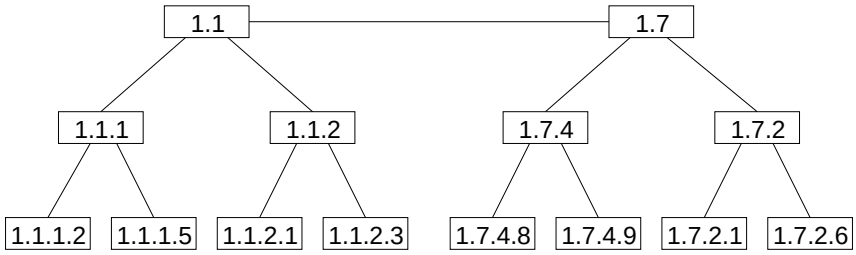


Abbildung 3.5.: Hierarchische Anordnung der Domänen bei CONGRESS.

Multiple Access) unabhängig von einem zugrundeliegenden Multicast-Protokoll eingesetzt werden. Beim Entwurf von CONGRESS sind folgende Richtlinien beachtet worden:

Kein Fluten: Bei Gruppenänderungen soll nicht das gesamte WAN über die Änderung informiert werden.

Hierarchisches Design: Die CONGRESS-Server sind in einer Baumstruktur angeordnet.

Robustheit: Netzwerkausfälle oder -rekonfigurationen können den temporären Ausfall von CONGRESS-Servern verursachen. Bei einer späteren Wiederherstellung der Verbindung soll das keinen Einfluss auf die Anwendungen haben.

CONGRESS betrachtet das Netzwerk als eine Hierarchie von Domänen, wobei jede Domäne von einem CONGRESS-Server bedient wird, wie in Abbildung 3.5 dargestellt. Die Adressen in Abbildung 3.5 sind die Domänenadressen und haben nichts mit den verwalteten IP-Adressen zu tun. Für jede IP-Multicast-Adresse (Gruppe) existiert ein aufspannender Teilbaum innerhalb der Hierarchie. In diesem Teilbaum sind alle Multicast-Router (Blätter im Baum) enthalten, die an der jeweiligen Gruppe partizipieren. Jeder CONGRESS-Server braucht nur die benachbarten Server (Vater- und Sohnknoten) im Baum zu kennen und an diese die jeweiligen Gruppenänderungen weiterzugeben. Hierdurch ist eine einfache Verwaltung im CONGRESS-Server möglich und die Teilbäume werden durch Join/Leave-Nachrichten der Multicast-Router konstruiert, bzw. modifiziert. Der Nachteil dieses Verfahrens ist, dass keine Aggregation von Informationen in den Servern stattfinden kann. Jeder Blattknoten im Baum muss alle anderen Blattknoten kennen, die an derselben Gruppen partizipieren. Hierdurch sind die CONGRESS-Server (und auch das Netzwerk) in den höheren Domänen stärker belastet und der Signalisierungsverkehr innerhalb der gesamten Hierarchie steigt linear mit der Anzahl der Gruppen und deren Mitglieder an. Der CONGRESS-Ansatz skaliert also nur bedingt und ist somit nicht optimal für einen Einsatz in einem Backbone-Netz geeignet.

IP-SENATE

IP-SENATE ist die zweite Komponente von IMSS. Die Aufgaben von IP-SENATE sind: 1. die Übertragung von IP-Multicast-Datagrammen über Shortcut-Verbindungen, 2. die

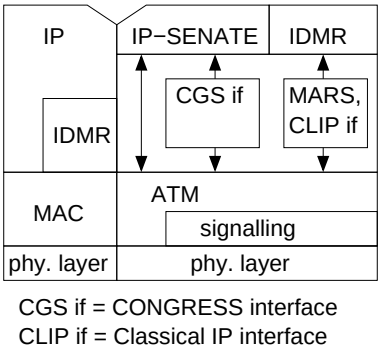


Abbildung 3.6.: Aufbau eines IP-SENATE Routers.

Einrichtung dieser Verbindungen und 3. Routing-Entscheidungen und die Kompatibilität zu IDMR-Protokollen. IP-SENATE ist nach folgenden Richtlinien entworfen worden:

- Best-Effort-Dienst: IP-SENATE garantiert keine Shortcut-Kommunikation, versucht aber diese immer zu ermöglichen.
- Shortcut-Kommunikation findet nur zwischen Routern und nicht direkt zwischen Endsystemen statt.
- IP-SENATE bietet als Kommunikationsschema VC Mesh, MCS und eine Hybridform aus beiden an.
- Migration von VC Mesh nach MCS und eine MCS-Lastverteilung ohne eine globale Rekonfiguration.
- IP-SENATE basiert auf CONGRESS für die Auflösung und Verwaltung der Multicast-Adressen.
- IP-SENATE ist inter-LIS basiert und erweitert nur die IDMR-Router. Für intra-LIS Kommunikation wird MARS benötigt.

Ein IP-SENATE-Router ist definiert als Border-Router, der eine Shortcut-Routing-Domäne mit einigen IDMR-Routingdomänen verbindet. IP-SENATE hat eine Schnittstelle zu CONGRESS, in der die Gruppen verwaltet werden, zum MARS und zu den IDMR-Protokollen (Abbildung 3.6). Im Folgenden werden die zwei wesentlichen Teile von IP-SENATE, das Shortcut-Routing und die IDMR-Schnittstelle, kurz vorgestellt.

Meldet sich ein Empfänger über IGMP oder MARS als Teilnehmer für eine Gruppe an, so werden mittels CONGRESS alle anderen beteiligten IP-SENATE-Router ermittelt. Dies ist auch bei einem Sender an eine Gruppe der Fall. Zu den ermittelten Routern wird dann entweder über eine Punkt-zu-Mehrpunkt-Verbindung eine ATM-Direktverbindung aufgebaut oder es wird eine Verbindung zu einem MCS (bei IMSS übernimmt ein IP-SENATE-Router ebenfalls die Funktionalität eines MCS) hergestellt, der die Daten an die zugehörigen Router der Gruppe weiterleitet. Bei Änderungen in der

Menge der zur Gruppe gehörigen Router müssen, analog zum MARS, entsprechende Änderungen an der Verbindung vorgenommen werden. Die Entscheidung, ob das VC-Mesh- oder MCS-Schema verwendet wird, hängt davon ab, ob sich ein IP-SENATE-Router dazu bereiterklärt, die MCS-Aufgabe für die betreffende Gruppe zu übernehmen. Ist das der Fall, wird das MCS-Schema angewendet, ansonsten das VC-Mesh-Schema.

Empfängt ein IP-SENATE-Router über die IDMR-Schnittstelle ein Datagramm, so wird dieses an alle anderen relevanten IP-SENATE-Router über eine Shortcut-Verbindung weitergeleitet. Relevant sind in diesem Zusammenhang alle IP-SENATE-Router, die sich ebenfalls für dieselbe Gruppe angemeldet haben. Ein IP-SENATE-Router kann Datagramme über die IDMR-Schnittstelle oder über Shortcut-Verbindungen empfangen. Hierdurch kann es passieren, dass ein IP-SENATE-Router über beide Schnittstellen dieselben Datagramme empfängt. Um diese Redundanz zu vermeiden, wird in so einem Fall die ausgehende IDMR-Schnittstelle abgeschnitten und es werden nur noch Datagramme von der Shortcut-Verbindung akzeptiert.

Bewertung

Das Konzept von IMSS behandelt die Integration einer IP-Multicast-Unterstützung im ATM-Backbone. Damit unterscheidet es sich von den bisher vorgestellten Konzepten, da bei IMSS keine direkte Anbindung von Endsystemen über ATM notwendig ist. Für die **Verwaltung** wird ein hierarchisches Schema eingesetzt. Durch diese verteilte **Organisation** wird eine erhöhte **Ausfallsicherheit** bei IMSS erreicht. Allerdings findet keine Aggregation der Teilnehmer in der Baumstruktur statt. Dadurch ist, analog zum MARS, ein hoher **Signalisierungsaufwand** bei Gruppenänderungen nötig.

Ein weiterer Schwerpunkt von IMSS ist die integrierte Kooperation mit **IDMR-Protokollen**. Für den **Datentransport** innerhalb des ATM-Netzes kann das MCS- oder VC-Mesh-Schema verwendet werden. Darüber hinaus unterstützt IMSS eine Kombination aus beiden Schemata, indem ausgewählte Quellen die Daten direkt über VC Mesh an die Empfänger verteilen und ansonsten das MCS-Schema angewendet wird. Hierbei sind allerdings die Kriterien unklar, wann genau welches Schema oder eine Kombination aus Beiden eingesetzt werden.

3.3. ATM-Schicht

Der wichtigste Punkt bei der Lösung der Gruppenkommunikation in der ATM-Schicht ist das Problem der Reihenfolge der Zellen bei AAL5 (siehe auch Unterkapitel 2.2.5 ab Seite 16). Hierdurch ist es nicht möglich, mehrere Verbindungen simultan auf eine einzige Verbindung zu multiplexen (Concast), da AAL5 keine zusätzlichen Informationen beinhaltet, zu welchem Datenpaket die Zellen des Datenpaketes gehören. Diese Information ist nur indirekt durch die sequenzielle Reihenfolge der Zellen vorhanden.

Bei den meisten der hier vorgestellten Ansätzen wird davon ausgegangen, dass ATM eine Mehrpunkt-zu-Mehrpunkt-Verbindung bereitstellt, allerdings ohne die oben erwähnte Problematik der Zellenreihenfolge dabei gelöst zu haben. Diese Voraussetzung stellt

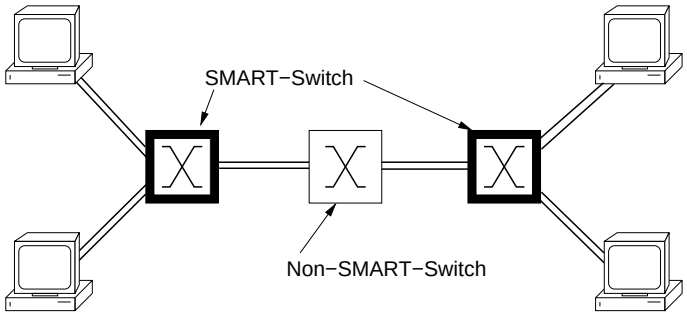


Abbildung 3.7.: SMART-Multicast-Baum mit zwei Mehrpunkt-zu-Mehrpunkt-Verbindungen.

zugleich auch das größte Hindernis für die im Folgenden vorgestellten Ansätze dar. Es sind Änderungen an der ATM-Schicht und evtl. höheren Schichten (AAL-Schicht, UNI-Signalisierung) notwendig. Aus praktischer Sicht sind daher die hier vorgestellten Ansätze kaum realisierbar, da im gesamten ATM-Netz Änderungen in den Komponenten (ATM-Schalteinheiten und ATM-Schnittstellenkarten) notwendig wären.

3.3.1. SMART

SMART (**S**hared **M**any-to-many **A**TM **R**eserva**T**ions) ist ein Protokoll, welches einen gemeinsam benutzten Multicast-Baum für die Gruppenkommunikation einsetzt. SMART unterstützt Verkehrsverträge mit expliziten Dienstgüten, benötigt keine Zellenpuffer und erhält die Zellenreihenfolge innerhalb des ATM-Netzes. SMART basiert grundlegend auf Mehrpunkt-zu-Mehrpunkt-Verbindungen, wobei Sender auch immer Empfänger sein müssen. Diese Verbindungen können aus Sicht einer ATM-Schalteinheit ohne großen Aufwand unterstützt werden [43]. Das dafür zu lösende Hauptproblem ist die Verschachtelung der Zellen, was insbesondere bei AAL5 sehr problematisch ist. Eine Lösung hierfür ist der Einsatz von AAL3/4, wobei aber wiederum eine koordinierte Verteilung von Multiplexing Identifiern (MID's) auf die Quellen vonnöten ist. Bei SMART wird diese Problematik umgangen. Des Weiteren ermöglicht SMART eine Verteilung auf Nachfrage (demand sharing) der reservierten Ressourcen auf alle Sender.

SMART basiert auf einer Token-Kontrolle, nur der im Besitz des Tokens befindliche Teilnehmer darf auf der ATM-Verbindung senden (ähnlich Token Ring). Für die Ver- und Weitergabe des Tokens zwischen den Sendern setzt SMART ein Protokoll mit Request- und Grant-Nachrichten ein. Ein Sender kann das Token anfordern und erhält es entweder umgehend, wenn kein anderer Sender im Besitz des Tokens ist, oder das Token wird vom aktuellen an den neuen Sender weitergegeben. Für die Anforderung und Weitergabe eines Tokens werden RM (Resource Management) Zellen eingesetzt, die immer an das Ende jedes Datenpaketes angehängt werden. Hierdurch werden das Netzwerk und die beteiligten Endsysteme immer über den aktuellen Zustand des Tokens informiert.

Pro Kommunikationsgruppe kann SMART mehrere parallele Mehrpunkt-zu-Mehrpunkt-Verbindungen (Abbildung 3.7) mit unterschiedlichen Dienstgüten anbieten. Da

pro Verbindung ein Token verwaltet wird, können dadurch auch mehrere Sender auf den parallelen Verbindungen gleichzeitig senden. In Kombination mit AAL3/4 und der Verwendung der von AAL3/4 zur Verfügung gestellten MID's (**M**ultiplexing **ID**entifier) kann die Anzahl der gleichzeitig aktiven Sender weiter erhöht werden.

Bewertung

Die an SMART beteiligten ATM-Schalteinheiten müssen eine entsprechend modifizierte ATM-Schicht besitzen. Über die Konfiguration von Virtual Paths können aber auch nicht modifizierte ATM-Schalteinheiten überbrückt werden. Des Weiteren basiert SMART auf Mehrpunkt-zu-Mehrpunkt-Verbindungen, die aber nicht genauer spezifiziert worden sind. Die Vergabe der MID's, um mehrere Sender pro Verbindung mit AAL3/4 zu ermöglichen, ist ebenfalls nicht geklärt. Da immer nur ein Sender aktiv sein kann, ist der **Ressourcenbedarf** von SMART im ATM-Netz im Verhältnis zur Datenmenge sehr hoch, dagegen kommt es aber zu keinerlei **Verkehrskonzentrationen**. Dies gilt insbesondere in Kombination mit Reservierungen für eine **Dienstgüteunterstützung**. Durch die Verwendung eines Token-Mechanismus zur Synchronisation der Sender ist SMART für viele Gruppenkommunikationsanwendungen nicht geeignet, da hierbei eine zu große Sendeverzögerung entsteht. Der **Signalisierungsaufwand** pro Teilnehmeranmeldung ist sehr gering, hingegen erfordert die Tokenvergabe eine kontinuierliche Signalisierung begleitend zum Datentransport.

3.3.2. SEAM

SEAM (**S**calable and **E**fficient **A**TM **M**ulticast) versteht sich als eine Weiterentwicklung der Punkt-zu-Mehrpunkt- zu Mehrpunkt-zu-Mehrpunkt-Verbindungen von ATM. Die Entwicklungsziele von SEAM sind Skalierbarkeit bzgl. der Netzwerkgröße, der Gruppengröße, deren Zusammensetzung und der Häufigkeit der Gruppenänderungen zu erreichen. Eine Gruppe wird durch eine Gruppenadresse identifiziert und alle Mitglieder (Sender und Empfänger) müssen sich explizit bei der Gruppe an- und abmelden.

SEAM basiert auf ein gemeinsam von allen Sendern und Empfängern einer Gruppe benutzten Baum, welcher durch eine einzige ATM-Verbindung repräsentiert wird. Für den Baum benutzt SEAM das Core-Based-Tree-Schema (CBT, Abbildung 3.8) [44, 45]. CBT's erleichtern die Signalisierung, und die Datenweiterleitung in einem Core ist genauso wie in jedem anderen Baumknoten. Der Core sollte für SEAM eine ATM-Schalteinheit sein. Nachteile von CBT's, wie etwas höhere Verzögerungen als bei Quellbäumen und Verkehrskonzentrationen werden dabei in Kauf genommen. Durch die Aufteilung verschiedener Gruppen auf mehrere Cores kann eine Verkehrskonzentrationen zum Teil vermieden werden, konkret hängt das aber von der Datenmenge ab, die die jeweiligen Gruppen produzieren.

SEAM benutzt ein Schema ähnlich dem Reverse Path Forwarding (RPF [46]) um im gemeinsamen Baum ein Shortcut-Routing zu ermöglichen (Abbildung 3.8). Bei einem Knoten werden ankommende Zellen nicht an alle ausgehenden Verbindungen weitergeleitet (wie bei RPF), sondern nur an die Verbindungen, wo auch Empfänger für

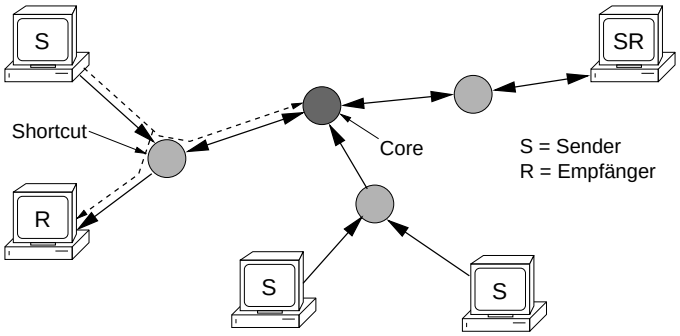


Abbildung 3.8.: Core Based Tree bei SEAM und Shortcut Routing.

diese Gruppe registriert sind. Damit wird erreicht, dass die Verkehrsmenge einer Gruppe bei jedem Knoten des Baumes dieselbe ist und Paketverzögerungen reduziert werden. Für dieses Verfahren ist es allerdings notwendig, dass in jedem Knoten für jeden Ausgangsport und Gruppe der Sender/Empfänger-Zustand gespeichert werden muss. SEAM verwendet AAL5, um die Pakete in Zellen zu segmentieren. Um die Verschachtelung von Zellen in einer ATM-Schalteinheit zu vermeiden, werden Puffer eingesetzt. In einer ATM-Schalteinheit wird auf einer Verbindung (Baum) zu einem Zeitpunkt immer nur ein AAL5-Zellenstrom weitergeleitet. Andere ankommende Zellenströme werden zwischengespeichert und dann nach dem Round-Robin-Verfahren abgearbeitet.

Bewertung

Bei SEAM ist nicht angegeben, wie und wann eine Core-Schalteinheit für eine Gruppe gewählt wird und ob jede SEAM-Schalteinheit in der Lage ist, eine Core-Schalteinheit zu sein. Des Weiteren muss bei allen ATM-Schalteinheiten im Baum das SEAM-Protokoll implementiert sein, nur bei Randknoten ist ein Tunneln über nicht SEAM-fähige Schalteinheiten möglich. Die Effizienz von SEAM hängt außerdem stark von der Anzahl der Sender und deren Verkehrsmenge ab, da es hier bei steigender Anzahl zu Pufferüberläufen und erhöhten **Verzögerungen** in einer SEAM-Schalteinheit kommen kann. Darüber hinaus erlaubt SEAM keinerlei **Dienstgüteunterstützung** beim Verbindungsaufbau.

Die **Organisation** der **Verwaltung** bei SEAM ist verteilt und die Gruppenteilnehmer werden in den Baumknoten aggregiert. Dadurch wird der **Signalisierungsaufwand** bei Teilnehmeran- und -abmeldungen stark reduziert. Die **Ausfallsicherheit** bei SEAM variiert. Je näher eine ausgefallene Komponente dem Core ist, desto mehr Systeme sind vom Ausfall betroffen.

3.3.3. SPAM

Eine Ergänzung zu SEAM stellt SPAM (**S**imple **P**rotocol for **A**TM **M**ulticast) [33] dar. SPAM behandelt die Problematik der Verschachtelung von Zellen, wenn mehrere Senderströme auf ein Verbindung gemultiplext werden (Concast). Bei SEAM werden AAL5-Pakete in der ATM-Schalteinheit gepuffert, wenn bereits ein anderes AAL5-Paket

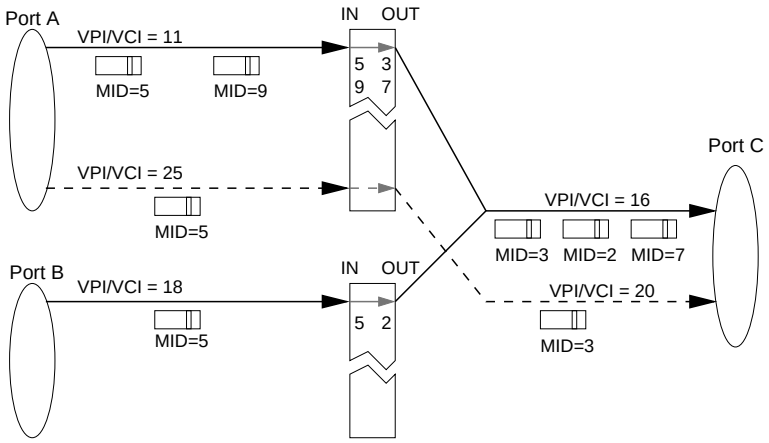


Abbildung 3.9.: Umordnung der MID's bei SPAM: Pro Eingangsport existiert eine Umordnungstabelle, die von allen ankommenden Verbindungen genutzt wird.

auf dieselbe Ausgangsleitung weitergeleitet wird. Das führt zu zwei Problemen: Erstens kann ein AAL5-Paket auf einer langsamen Datenleitung eine andere schnellere Datenleitung erheblich verzögern, da erst das Ende des langsamen Paketes abgewartet werden muss und zweitens resultiert der Verlust der letzten Zelle eines AAL5-Paketes in den zusätzlichen Verlust des darauffolgenden Paketes.

Bei SPAM wird ein Weiterleitungsmodell vorgeschlagen, das die obigen Probleme beseitigen kann. SPAM verwendet Senderkennungen, wie den MID (**M**ultiplexing **ID**entifier) bei AAL3/4, um die Zellen der verschiedenen Sender unterscheiden zu können und somit ein unverzügliches Weiterleiten der Zellen in den Schalteinheiten zu ermöglichen. Da AAL3/4 aber einen erheblichen Overhead pro Zelle transportiert (44 Byte Nutzdaten, 4 Byte AAL3/4-Kopf pro Zelle) und im Gegensatz AAL5 keinen weiteren Overhead pro Zelle hat, wird ein neues AAL-Format als Kompromiss zwischen AAL3/4 und AAL5 vorgeschlagen: AAL-SPAM. AAL-SPAM ist eine Erweiterung von AAL5 mit einer 15 Bit MID Kennung und 1 Bit für die Markierung eines Paketes als Paketanfang (Beginning of Packet, BOP) oder Paketfortsetzung (Continuation of Packet, COP). Das Paketende (End of Paket, EOP) wird wie bei AAL5 im Payload-Feld des Zellenkopfs gespeichert.

Ein Grundproblem bei der Vergabe von MID's ist, dass diese Vergabe konfliktfrei sein muss, da ansonsten verschiedene Sender nicht mehr unterschieden werden können. Bei SPAM wird diese Problematik durch einen Algorithmus (Smart MID Remapping Algorithm) gelöst, der die MID's der Pakete dynamisch umordnet (Abbildung 3.9). Die Umordnung basiert auf einem zirkulären Puffer der Größe 2^{15}, in dem freie MID's gespeichert werden.

Bewertung

SPAM stellt keinen eigenständigen Ansatz zur Problematik der Gruppenkommunikation und ATM dar. Durch SPAM kann die **Verzögerung** in den Knoten minimiert werden, da ein Zwischenspeichern der Datenpakete in den Knoten entfällt. Das wird mit der

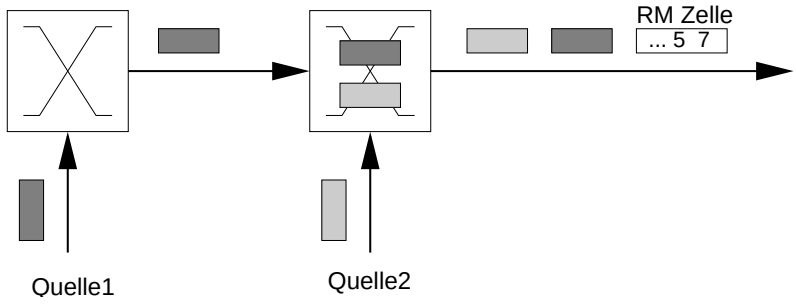


Abbildung 3.10.: CRAM sendet eine RM-Zelle vor den Zellen verschiedener Sender, die deren SID's transportiert.

Einführung eines neuen **Datenformates** in der AAL-Schicht erreicht. Hierzu sind wiederum Änderungen in den angeschlossenen Endsystemen notwendig. Eine zusätzliche Bearbeitung dieser AAL-Daten in den ATM-Schalteinheiten ist ebenso erforderlich, wodurch auch in den Zwischensystemen Änderungen unumgänglich für den Einsatz von SPAM sind.

3.3.4. CRAM

CRAM (**C**ell **R**e-labeling **A**t **M**erge-points) [34] verfolgt dasselbe Ziel wie SPAM: Eine integrierte Zellenweiterleitung für verteilte Multicast-Bäume. Um das Problem der Verschachtelung von Zellen beim Multiplexen mehrerer Quellen zu vermeiden, setzt SPAM ein neues AAL-Format ein, um damit effizient MID's zu speichern. In CRAM wird hingegen die Einführung eines neuen AAL-Formats umgangen, und zwar in dem externe Informationen (MID's) in zusätzlichen RM-Zellen (Resource Management, definiert im Zusammenhang mit dem ABR-Service) transportiert werden.

In CRAM werden sogenannte SID's (**S**ource **ID**entifier) verwendet, die ein Äquivalent zu MID's darstellen. Die Vergabe von SID's kann dabei entweder global erfolgen oder es kann wie bei SPAM ein Umordnungsalgorithmus eingesetzt werden. Jede Quelle muss an den Anfang ihres Netzwerkschichtpaketes ihre SID schreiben. In den ATM-Schalteinheiten hat die Logik der Eingangspuffer dann dafür zu sorgen, dass jede zu einem Paket gehörige Zelle diese SID zugeordnet bekommt. Treffen jetzt an einer ATM-Schalteinheit Datenpakete (Zellenströme) von mehreren Quellen aufeinander, so wird vor den gemultiplexten ausgehenden Zellen eine RM-Zelle gesendet, die die SID-Nummern der nachfolgenden Zellen enthält. Das wird im Konzept als RM cell train bezeichnet (Abbildung 3.10). Bei einer SID Größe von zwei Byte kann eine RM-Zelle für maximal 24 Datenzellen deren MID's transportieren. Dann muss eine neue RM-Zelle generiert werden. Ist kein Multiplexing notwendig, also treffen nur Zellen von einem Sender ein, so werden auch keine RM-Zellen generiert.

Ein weiteres Detail ist hier noch zu beachten, und zwar wann der Block mit der RM-Zelle und den Datenzellen gesendet werden soll. Wenn eine noch fehlende Zelle für einen Block verspätet eintrifft, kann sich dadurch das Versenden des ganzen Blockes verzögern. Um diese Verzögerung zu begrenzen, wird ein Timeout-Mechanismus eingesetzt, der eine

	SPAM	CRAM
Datenformat	AAL-SPAM	AAL5 + Anwendung
Verzögerung	minimal	erhöht
Robustheit	normal	empfindlich

Tabelle 3.2.: Vergleich von SPAM und CRAM.

maximale Sendeverzögerung garantiert.

Bewertung

Das größte Problem bei CRAM ist der Verlust von Zellen, insbesondere von RM-Zellen. Hierdurch wird es auf Empfängerseite unmöglich, die nachfolgenden Zellen zu reassemblieren und alle davon betroffenen Pakete müssen verworfen werden. Das mindert die **Robustheit** von CRAM bezüglich Zellverlusten. Des Weiteren ist ein zusätzliches Zwischenspeichern von Zellen notwendig, da eine RM-Zelle vor den zugehörigen Datenzellen gesendet wird, in der RM-Zelle aber die SID's der nachfolgenden Datenzellen enthalten sein müssen. Hierdurch entsteht eine höhere **Verzögerung** als bei SPAM.

3.4. Zusammenfassung

Die hier vorgestellten Arbeiten werden in diesem Unterkapitel nach den in Unterkapitel 3.1 festgelegten Kriterien miteinander verglichen. Zusätzlich wird der in dieser Arbeit vorgestellte Ansatz (SkaGAN, siehe Kapitel 4, 5 und 6) in den Vergleich mit aufgenommen, um eine allgemeine Einschätzung zu den bestehenden Arbeiten zu ermöglichen.

Bei den in diesem Kapitel vorgestellten Arbeiten nehmen die Ansätze SPAM und CRAM (Unterkapitel 3.3.3 und 3.3.4) eine Sonderstellung ein. Diese beiden Ansätze behandeln nur eine Teilproblematik, und stellen keine vollwertigen Lösungen im Bereich Gruppenkommunikation über ATM dar. Hieraus ergibt sich, dass nur ein Teil der vorgestellten Kriterien auf diese beiden Ansätze anwendbar sind. Daher werden SPAM und CRAM in Tabelle 3.2 gesondert zusammengefasst und nur nach den Kriterien Datenformat, Verzögerung und Robustheit beurteilt. Alle anderen Ansätze und SkaGAN sind in Tabelle 3.3 zusammengefasst.

Der Hauptnachteil der in Tabelle 3.2 aufgeführten Ansätze SPAM und CRAM ist das Datenformat. Beide Ansätze verwenden ein modifiziertes AAL5-Format. Bei SPAM ist hierzu eine Änderung in der ATM-Schicht notwendig und bei CRAM muss der Ansatz von der darüber liegenden Anwendung unterstützt werden. Diese Voraussetzungen bei SPAM und CRAM lassen die Durchführbarkeit und Realisierbarkeit beider Ansätze fragwürdig erscheinen, da zu viele Änderungen an bestehenden Systemen und Programmen nötig sind.

Wie aus dem Vergleich in Tabelle 3.3 ersichtlich ist, basieren die meisten Ansätze im Bereich Gruppenkommunikation über ATM entweder auf dem VC-Mesh- oder auf dem

	MARS		VENUS	EARTH	verteilter MARS	mehrere MCS	PIM-SM	IMSS	SMART	SEAM	SkaGAN lokal	SkaGAN global
Datentransport												
Schema	VC Mesh	MCS	VC Mesh o. MCS	VC Mesh	VC Mesh o. MCS	mehrere MCS	VC Mesh	VC Mesh o. MCS	Baum[1]	CBT	mehrere MCS	Baum
Verkehrskonzentration	verteilt	im MCS	[2]	[2]	[2]	verteilt auf MCS	[2]	[2]	nein	im Core	verteilt auf MCS	
Verzögerung	gering	moderat[3]	[2]	[2]	[2]	moderat	[2]	[2]	gering	moderat	moderat	moderat-erhöht
Datenformat	unverändert	Einkapselung	[2]	[2]	[2]	[2]	[2]	[2]	AAL5, AAL3/4	AAL5	Einkapselung	
Dienstgüteunterstützung	nein	nein	nein	QoS-Ebenen	nein	nein	nein	nein	ja	nein	nein	nein
Fehlertoleranz	gut	SPoF[4]	[2]	[2]	[2]	hoch	[2]	[2]	hoch	normal	hoch	hoch
Ressourcenbedarf	$(N-1)^2$	$2N + (N+1)$	[2]	[2]	[2]	$2N + L(N+1)$	[2]	[2]			$2N + L(N+1)$	$2N + N\frac{k+1}{k} + 2\log_k N$[6]
Verwaltung												
Organisation	zentral	zentral	k.A.	zentral	verteilt	zentral	verteilt, aggreg.	verteilt	verteilt	verteilt, aggreg.	zentral	verteilt, aggreg.
Ausfallsicherheit	SPoF[4]	SPoF[4]	k.A.	Backup-Systeme	Backup-Systeme	SPoF[4]	nein	ja	Teilbaum	Teilbaumausfall	SPoF[4]	Teilbaumausfall
Signalisierungsaufwand	N	MARS + MCS	k.A.	[2]	[2]	MARS + L * MCS	[2]	[2]	1[5]	1	MARS + L * MCS	$\log_k N$[6]
IDMR-Protokolle	nein	nein	nein	ja	nein	nein	ja	ja	nein	nein	nein	nein

Tabelle 3.3.: Vergleich der Konzepte zur Gruppenkommunikation über ATM.

[1] Keine genaueren Angaben zur Baumstruktur, bzw. -konstruktion.
[2] Der Wert ist abhängig vom verwendeten Schema (MCS oder VC Mesh), Angaben siehe Spalten bei MARS.
[3] Die Verzögerung ist von der Anzahl der Sender abhängig.
[4] **S**ingle-**P**oint-**o**f-**F**ailure.
[5] Tokenvergabe erfordert regelmäßigen Nachrichtenaustausch, auch ohne Änderung der Gruppenteilnehmer.
[6] Unter der Annahme, dass ein annähernd ausgeglichener Baum vorliegt. Mehr im Text.

MCS-Schema. Eine Ausnahme bilden SMART und SEAM, die auf einer Baumstruktur aufbauen. Diese Ansätze nehmen aber auch eine Sonderstellung ein, da sie in der ATM-Schicht realisiert werden müssen. Der in den folgenden Kapiteln vorgestellte Ansatz ist eine Kombination zwischen MCS- und Baum-Schema und ist in der Anwendungsschicht realisiert. Eine Realisierung in der Anwendungsschicht hat den Vorteil, dass keinerlei Änderungen an bestehenden ATM-Systemen nötig sind, um den Ansatz einzusetzen. Darüber hinaus ist immer noch eine spätere Integrierung von Teilen des Ansatzes in der ATM-Schicht möglich.

Die Verkehrskonzentration ist ein inhärentes Problem beim MCS-Schema und kann durch den Einsatz mehrerer MCS verringert werden. Hingegen sind beim MCS-Schema immer die Paketreflexionen zu beachten, was durch Einkapselung der Anwendungsdaten geschieht. Die Beurteilung der Verzögerung geht mit der Verkehrskonzentration einher. Die Einschätzung der Verzögerung hat als Maßstab die Größenordnung der Zellverzögerung in der ATM-Schicht (im Bereich 10^{-10} Sekunden). Das führt dazu, dass das MCS-Schema schon als moderat bezüglich der Verzögerung eingeschätzt wird. Beim MCS-Schema findet im MCS eine zusätzliche Reassemblierung der Datenpakete statt, die zu einer deutlich messbaren Erhöhung der Verzögerung führt.

Eine angemessene Dienstgüteunterstützung [47] lassen alle Ansätze vermissen, obwohl ATM hierzu alle Voraussetzungen anbietet. Ein Hauptgrund hierfür dürfte die Aggregation von sendenden Gruppenteilnehmern sein, die zu einem nicht vorhersehbaren und veränderlichen Datenaufkommen führt. Für das veränderliche Datenaufkommen kann mit ATM kein optimaler Verkehrsvertrag abgeschlossen werden. Die Fehlertoleranz der Ansätze stellt hingegen eine andere Art der Dienstgüte dar. Die Robustheit eines Ansatzes wird als hoch eingestuft, wenn der Ausfall von Systemen nur deren lokalen Bereich betrifft und die anderen Systeme (evtl. nach kurzer Unterbrechung) weiter Daten austauschen können. Der in dieser Arbeit vorgestellte Ansatz weist diesbezüglich eine hohe Robustheit auf, da ausgefallene oder nicht mehr erreichbare Systeme durch andere Systeme ersetzt werden können.

Das Kriterium Ressourcenbedarf bietet ein Maß für den logischen Ressourcenverbrauch anhand der benötigten ATM-Verbindungsendpunkte. Hier wird nicht das Datenaufkommen im ATM-Netzwerk beurteilt, das zu den hier vorgestellten Maßen keine Beziehung haben muss. Bei der Bewertung von SkaGAN ist zusätzlich zur Gruppengröße N noch der Faktor k hinzugekommen. Dieser Faktor ist ein Maß für die Aufteilung der Gruppenmitglieder auf lokale Teilgruppen, bei $k = N$ entspricht das dem MCS-Schema. Hier ist ein erhöhter Ressourcenbedarf gegenüber den anderen Ansätzen festzustellen. Dieser Bedarf ist durch die für die Baumstruktur zusätzlich benötigten Verbindungen zu erklären.

Die Organisation der Verwaltung ist bei den meisten Ansätzen zentral, was in der Regel auch beim MCS- oder VC-Mesh-Schema Vereinfachungen bei der Realisierung ermöglicht. Zentrale Verwaltungen skalieren allerdings nicht mit der Gruppengröße und es kann hierdurch zu zusätzlichen Verzögerungen bei der An- und Abmeldung kommen. Ein anderer Faktor ist die Ausfallsicherheit, die bei zentralen Systemen im Allgemeinen als ungenügend angesehen werden kann. Verbesserungen bringen hier Backup-Maßnahmen der zentralen Verwaltungssysteme. Ansonsten ist die Ausfallsicherheit eng gekoppelt

an die Robustheit. Beide Aspekte sind zu beachten, damit die Gruppenkommunikation beim Ausfall von Komponenten stabil arbeitet.

Der Signalisierungsaufwand ist ein Maß, welches die ungefähr benötigte Nachrichtenanzahl pro Gruppenteilnehmeranmeldung repräsentiert. Dieses Maß sollte im Idealfall wesentlich kleiner als die Gruppengröße N sein oder davon unabhängig. Nur in diesem Fall kann der Ansatz auch für große Gruppen eingesetzt werden.

Das in dieser Arbeit entwickelte Konzept hat das Ziel, Probleme der Skalierbarkeit bei der Gruppenkommunikation über ATM zu lösen. Die untersuchten Forschungsarbeiten zeigen hier alle ähnliche Mängel im Bereich Datentransport und Gruppenverwaltung. Die Ursache liegt dabei im Wesentlichen in der Verwendung des VC-Mesh- oder MCS-Schemas, die aber nur für eine begrenzte Anzahl Endsysteme geeignet sind. Hier wird SkaGAN für den Weitverkehrsbereich ein neues Schema vorstellen, das die Gruppenteilnehmer in einer Baumstruktur anordnet und somit eine weitgehend lokale Verwaltung der Teilnehmer ermöglicht (Kapitel 6).

Im Bereich lokaler ATM-Netze zeigt sich, dass das VC Mesh und MCS Schema ebenfalls nicht in der Lage sind, alle Anforderungen zu erfüllen. Bei SkaGAN wird für die lokalen ATM-Netze eine Variante des MCS-Schemas vorgeschlagen, das mehrere MCS in Kombination mit einer Lastverteilung erlaubt (Kapitel 5).

4. SkaGAN: Überblick und Komponenten

In diesem und den beiden folgenden Kapiteln wird der Ansatz für eine **Ska**lierbare **G**ruppenkommunikationsunterstützung für **A**TM-**N**etze (SkaGAN) vorgestellt. Der Ansatz besteht dabei im Wesentlichen aus zwei Teilen:

Lokal: Gruppenkommunikationsunterstützung für lokale ATM-Netze und

Global: Verwaltung und Datentransfer für Gruppen in ATM-Weitverkehrsnetzen.

Zunächst werden jedoch die gemeinsam benutzten Komponenten und Protokolle erläutert, die in beiden Teilen des Ansatzes verwendet werden. Hierzu gehört das Verbindungsmanagement, der Nachrichtenaustausch und allgemeine Konventionen zum Nachrichtenformat. Der Zusammenhang zwischen den einzelnen Teilen des Ansatzes ist in Abbildung 4.1 dargestellt. Die gestrichelt gezeichneten Teile sind allgemeiner Natur und bilden die Grundlage, auf der die Konzepte für die lokale und globale Gruppenkommunikationsunterstützung aufbauen. Basierend auf dem MARS-Ansatz (Kapitel 2.4, ab Seite 23) wird ein Konzept zur Unterstützung von mehreren MCS vorgestellt, das eine bessere Skalierbarkeit und Lastverteilung in lokalen ATM-Netzen ermöglicht. Das Konzept für ATM-Weitverkehrsnetze ist in mehrere Bestandteile untergliedert. Die Basis bildet ein Schema für die Gruppenverwaltung, mit dem die Gruppenteilnehmer im ATM-Netz lokalisiert werden können. Auf diesen Informationen baut das Konzept für den Datentransfer zwischen den Gruppenteilnehmern auf. Dieses Basiskonzept wird anschließend noch erweitert, so dass auch der zuerst vorgestellte lokale Ansatz in das Gesamtkonzept für SkaGAN einfließen kann.

In Unterkapitel 4.1 werden zuerst die an der Gruppenkommunikation beteiligten Komponenten beschrieben. Die von allen Komponenten gemeinsam verwendeten Mechanismen und allgemeine Konventionen, wie Nachrichtenformate oder der Nachrichtentransport, werden in Unterkapitel 4.2 spezifiziert.

Auf diesem Kapitel bauen die nächsten beiden Kapitel 5 und 6 auf, die die Ansätze für eine Gruppenkommunikationsunterstützung in lokalen Netzen und Weitverkehrsnetzen beschreiben.

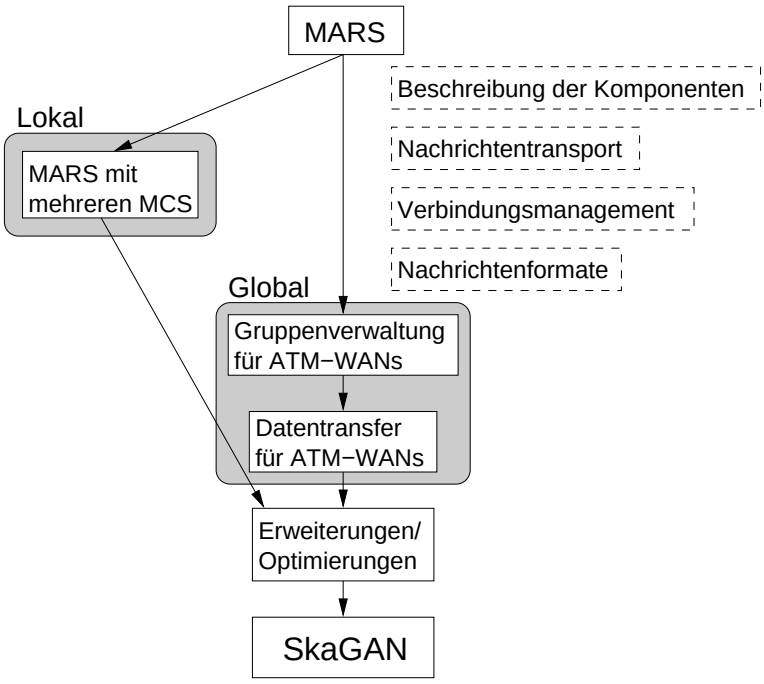


Abbildung 4.1.: Zusammenhang der einzelnen Teilkonzepte.

4.1. Komponenten

Bei SkaGAN können zur Unterstützung der Gruppenkommunikation in ATM-Netzen drei verschiedene Typen von Komponenten, bzw. Systemen unterschieden werden:

Endsystem: (Host) Repräsentiert einen normalen Rechner mit einer oder mehreren Anwendungen die die Gruppenkommunikation nutzen. Für das Konzept von SkaGAN ist es dabei nur wichtig, dass mehrere Gruppenkommunikationsverbindungen pro Endsystem existieren können. Vereinfachend kann ein Endsystem somit auch einen Multicast-Router repräsentieren, der eine Verbindung zu einem IP-Subnetz herstellt. Der Unterschied ist dabei für SkaGAN nur prinzipieller Natur. Während ein normales Endsystem höchstwahrscheinlich an eher wenigen Gruppen teilnimmt, partizipiert ein Multicast-Router an allen im IP-Subnetz vorkommenden Gruppen. Die Unterscheidung zwischen Multicast-Router und Endsystem ist daher höchstens an der Anzahl aktiver Gruppen und dem damit zusammenhängenden Datenverkehr zu treffen.

Multicast Server: (MCS) Der Multicast Server stellt von der Funktionalität her einen Proxy-Server dar, dessen Aufgabe im Wesentlichen in der Datenweiterleitung besteht. Der MCS leitet dabei Daten von Endsystemen oder anderen MCS weiter, und zwar wiederum an Endsysteme und andere MCS. Die wichtigste Eigenschaft des MCS ist dabei die Aggregation von Datenströmen, so dass mehrere ankommende Verbindungen auf eine ausgehende Verbindung gemultiplext werden. Hierdurch wird eine Zusammenfassung mehrerer Endsysteme ermöglicht, die durch

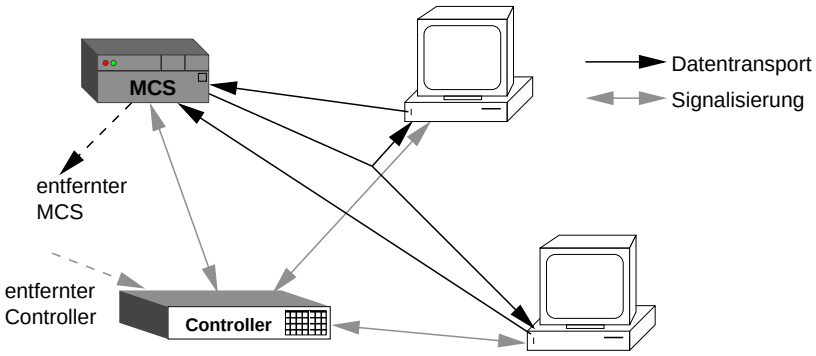


Abbildung 4.2.: Komponenten der Gruppenkommunikation und deren Signalisierungsverbindungen.

einen MCS nach außen hin repräsentiert werden.

Ein MCS ist eine Komponente in der Anwendungsschicht. Die MCS-Komponente kann dadurch in einem Endsystem, einem speziellen Server oder auch innerhalb einer ATM-Schalteinheit untergebracht sein.

Controller: (CTRL) Im Controller ist die Steuerungslogik für die Etablierung einer Gruppenkommunikation untergebracht. Der Controller ist die Komponente, in der sämtliche (Signalisierungs-) Nachrichten von Endsystemen, MCS oder anderen Controllern ausgewertet und entsprechende Reaktionen (Versenden weiterer Nachrichten) koordiniert werden. Auslösende Ereignisse sind Bei- und Austritte eines Endsystems zu/von einer Gruppe. Ein Controller kann ebenfalls wie ein MCS in einem Endsystem, einer ATM-Schalteinheit oder auch parallel zu einem MCS als Anwendungsprogramm arbeiten.

Das Zusammenspiel der drei Komponenten ist in Abbildung 4.2 dargestellt. Signalisierungsverbindungen existieren zwischen Endsystem/MCS und Controller und zwischen Controllern (gestrichelte Verbindung). Es gibt keine Signalisierungsverbindungen zwischen Endsystem und MCS. Datentransportverbindungen sind hingegen nur zwischen Endsystemen und MCS vorhanden, der Controller transportiert keine Anwendungsdaten. Die Signalisierungsverbindungen sind bidirektionale Punkt-zu-Punkt-Verbindungen, die Datentransportverbindungen sind hingegen unidirektional und können entweder Punkt-zu-Punkt- oder Punkt-zu-Mehrpunkt-Verbindungen sein.

4.2. Grundlegende Mechanismen und Definitionen

Für die im Folgenden erläuterten Verfahren wird vorausgesetzt, dass alle bei diesem Ansatz beteiligten Komponenten eine ATM-Schnittstelle haben und an einem ATM-Netzwerk angeschlossen sind. Diese Voraussetzung ist bei allen Komponenten identisch und die ATM-Schnittstelle, die den Zugriff für Anwendungen auf die ATM-Dienste ermöglicht, ist ebenfalls einheitlich. Die ATM-Schnittstelle besteht dabei im Kern aus zwei Teilen, der Verbindungskontrolle und dem Datentransfer.

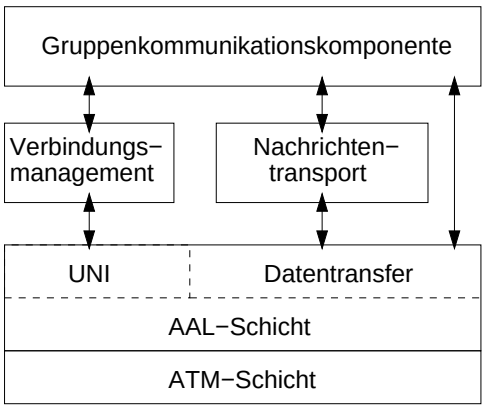


Abbildung 4.3.: Die ATM-Schnittstelle und die bei allen Komponenten verwendeten Module.

Der gesamte Datentransfer geschieht über SVC's (Unterkapitel 2.2.2, Seite 13) mit AAL5 (Unterkapitel 2.2.5, Seite 16). Um die Art der transportierten Daten unterscheiden zu können werden folgende Namenskonventionen verwendet (siehe auch Anhang E):

Paket AAL5-Paket.

Datenpaket Paket einer Anwendung

Nachrichtenpaket/Nachricht Paket der Signalisierung von SkaGAN

Abbildung 4.3 zeigt die einzelnen Module einer Komponente. Das Verbindungsmanagement (Unterkapitel 4.2.3) ist für den Auf- und Abbau von ATM-Verbindungen zuständig, unabhängig von deren späterer Verwendung. Das Modul für den Nachrichtentransport (Unterkapitel 4.2.2) bietet einen zuverlässigen Transportdienst auf speziellen Signalisierungsverbindungen an, und zusätzlich besteht noch die Möglichkeit des direkten Datentransportes über ATM. Zunächst wird aber das Konzept für die verwendeten Nachrichtenformate beschrieben.

4.2.1. Nachrichtenformate

Der Begriff Nachrichten wird im weiteren Verlauf immer für die Signalisierung verwendet, im Gegensatz hierzu bezeichnen Daten oder Datenpakete immer Anwendungsdaten.

Die hier verwendeten Nachrichtenformate orientieren sich an den bei verschiedenen Protokollen, wie ITU-T UNI Signalisierung (also Q.2931, etc.), ATM Forum UNI und PNNI Signalisierung, benutzten Nachrichtenformaten [48]. Alle Nachrichten haben einen gemeinsamen Nachrichtenkopf (Tabelle 4.1), an den sich je nach Typ unterschiedliche Inhalte anschließen können. Die Protokoll ID enthält die Versionsnummer des verwendeten Protokolls. Die verschiedenen Nachrichten werden durch ihren Nachrichtentyp identifiziert, die Nachrichtenlänge gibt den Umfang des eigentlichen Nachrichteninhalts an.

Der Nachrichteninhalt wiederum besteht aus einer Menge von Inhaltsfeldern, den sogenannten Informationselementen (information elements, IE). Alle IEs haben das gleiche

Feld	Oktett
Protokoll ID	1
Nachrichtentyp	2-3
Nachrichtenlänge	4-5
variabler Nachrichteninhalt	6-

Tabelle 4.1.: Von allen Nachrichten verwendeter Kopf.

Feld	Oktett
Informationselement ID	1
Länge	2
Inhalt	3-

Tabelle 4.2.: Aufbau eines Informationselementes (IE).

Basisformat (Tabelle 4.2), welches dem Format für den Nachrichtenkopf sehr ähnlich ist. Die Informationselement ID gibt den Typ an, und die Länge bestimmt die Größe des Inhaltsbereiches.

Für die Informationselemente sind eine Reihe von Basis-IEs (Tabelle 4.3) definiert. Diese stellen eine Art Grundvokabular dar, aus dem weitere IEs zusammengesetzt werden können.

Als Beispiel für eine Nachricht, die aus IEs zusammengesetzt ist, zeigt Tabelle 4.4 das Nachrichtenformat für Bestätigungen, wie sie vom Nachrichtentransportmodul (Unterkapitel 4.2.2) verwendet werden. Mit den Informationen aus Tabelle 4.4 ist der Inhalt für eine Bestätigungs-Nachricht vollständig spezifiziert und das endgültige Nachrichtenpaket (Typ: **ACK**) sieht dann wie in Tabelle 4.5 aus.

Die Definitionen aller Nachrichten befinden sich im Anhang C ab Seite 177.

Name	Länge (Oktetts)	Beschreibung
IP_Address	4	Eine IP-Unicast- oder -Multicast-Adresse
ATM_Address	20	spezifiziert beim ATM Forum (E.164)
SAP	2	Service Access Point, identifiziert Dienst beim Empfänger
Number	2	Nummer $[0, \ldots, 2^{16} - 1]$
LongNumber	4	Nummer $[0, \ldots, 2^{32} - 1]$
Boolean	1	boolescher Wert, entweder 0 (false) oder 1 (true)
FixedFloat	4	Festkommazahl $[0, \ldots, 2^{16} - 1]$, Auflösung $1/65535 = 1.5259 * 10^{-5}$

Tabelle 4.3.: Basisinformationselemente.

ACK		
Name	IE	Beschreibung
Message_Type	Number	Typ der zu bestätigenden Nachricht
ID	Number	Identifikationsnummer der Nachricht

Tabelle 4.4.: Nachrichtenformat für Bestätigungen.

Feld		Oktett
version:1		1
Typ: ACK		2-3
Länge: 8		4-5
Inhalt:	Typ: Number	6
	Länge: 2	7
	Message_Type	8-9
	Typ: Number	10
	Länge: 2	11
	ID	12-13

Tabelle 4.5.: Der Aufbau eines Nachrichtenpaketes für Bestätigungen.

Das einzige Format, das von diesem Schema abweicht, ist das für den Transport von Anwendungsdatenpaketen. Hierbei kommt es auf eine schnelle und effiziente Bearbeitung an, und nicht auf Erweiterbarkeit und Modularität wie bei den Nachrichtenpaketen. Das Format der Datenpakete ist in Tabelle 4.6 dargestellt. Die Sender- und Gruppenadresse bestehen aus je vier Oktetts und entsprechen damit der Größe von IPv4-Adressen. IP-Multicast basierte Anwendungen können somit ohne weitere Adressenabbildungen unterstützt werden. ATM-SAP steht für ATM-**S**ervice **A**ccess **P**oint und identifiziert den Dienst bzw. die Anwendung beim Empfänger, die das Datenpaket annehmen soll. Das Labelfeld hat die gleiche Länge wie eine Senderadresse und dient zur Erkennung von Paketreflektionen (siehe auch Unterkapitel 6.2.1, ab Seite 6.2.1). Das Inhaltsfeld dient zur Aufnahme der eigentlichen Anwendungsdaten. Im Kopf des Datenpaketes sind keinerlei Längenangaben vorhanden, die Größe des Inhaltsfeldes wird aus der Größe des AAL5-Paketes minus dem Datenpaketkopf (14 Byte) bestimmt. Unter Berücksichtigung der maximalen AAL5-Paketgröße können somit höchstens $65535 - 14 = 65521$ Bytes Anwendungsdaten in einem Datenpaket transportiert werden.

Feld	Oktett
Senderadresse	1-4
ATM-SAP	5-6
Gruppenadresse	7-10
Label	11-14
Inhalt	15-

Tabelle 4.6.: Kopf für Pakete mit Anwendungsdaten.

4.2.2. Nachrichtentransport

Alle an der Gruppenkommunikation beteiligten Komponenten verwenden für den Transport von (Signalisierungs-)Nachrichten ein Modul zur zuverlässigen Nachrichtenübertragung. Hierdurch kann die Kontrolllogik der Komponenten stark vereinfacht werden, da Fehlerfälle, wie z. B. verloren gegangene Nachrichtenpakete, nicht mehr berücksichtigt werden müssen.

Bei einer Integration von SkaGAN in die ATM-Schicht kann das hier beschriebene Protokoll durch das bestehende SAAL-Protokoll ersetzt werden. Auf den Einsatz eines bestehenden Transportprotokolls, wie TCP, wird verzichtet. Hierzu hätte TCP zuerst an ATM angepasst werden müssen [49]. Das betrifft zum einen die Adressierung (statt IP-Adressen und Portnummern, ATM-Adressen und SAP-Kennungen) und zum anderen die Auflösunfg von Abhängigkeiten (maximale Transfer Unit Größe, initialer Time-out) zwischen TCP und IP.

Das Modul für die Nachrichtenübertragung geht implizit immer davon aus, dass kein Netzwerk- oder Serverausfall vorliegt. Dies bedeutet, dass die Nachricht aus Sicht des Transportmoduls immer ausgeliefert werden kann. Diese Vereinfachung kann gemacht werden, da Ausfälle im Netzwerk oder beim Kommunikationspartner von der ATM- oder AAL-Schicht festgestellt werden, und die betroffene Verbindung daraufhin abgebaut wird. Dieser Fall wird vom Modul für das Verbindungsmanagement behandelt (Unterkapitel 4.2.3) und von dort der Gruppenkommunikationskomponente gemeldet.

Das Protokoll arbeitet ähnlich wie das Stop-and-Wait-Protokoll. Eine empfangene Nachricht wird mit einer positiven Bestätigung quittiert. Verloren gegangene Nachrichten werden nach einem Timeout auf Senderseite wiederholt und jede Nachricht erhält eine individuelle Sequenznummer. Diese Nummer wird bei der Bestätigung ebenfalls angegeben (siehe auch vorhergehendes Unterkapitel 4.2.1, Nachrichtenformat für Bestätigung, ID-Feld). Damit können auf Empfängerseite doppelt empfangene Nachrichten verworfen und auf Senderseite mehrfach eingetroffene Bestätigungen ignoriert werden.

Der Ablauf des Protokolls ist in Abbildung 4.4 dargestellt. Quelle und Senke sind dabei entweder Endsystem, MCS oder Controller. Die Nachricht wird bei der Quelle einmalig dem Nachrichtentransportmodul übergeben, das alle weiteren Aufgaben übernimmt. Auf der Empfängerseite übergibt das Transportmodul das angekommene Nachrichtenpaket der Senke. Der Anhang '[x]' ist die an das Nachrichtenpaket und die Be-

stätigung angehängte Sequenznummer, die im Zusammenhang mit der Senderadresse eine eindeutige Identifikation der Nachricht erlaubt. Das Nachrichtentransportmodul hat einen temporären Cache, in dem versendete oder empfangene Sequenznummern mit den zugehörigen Adressen gespeichert werden. Das ermöglicht das Wiedererkennen und Verwerfen von bereits empfangenen Datenpaketen und Bestätigungen. Im Cache werden die Informationen für eine Zeitdauer vom dreifachen Timeout-Wert gehalten. Abbildung 4.5 zeigt die zugehörigen Flussdiagramme für die Quelle und die Senke.

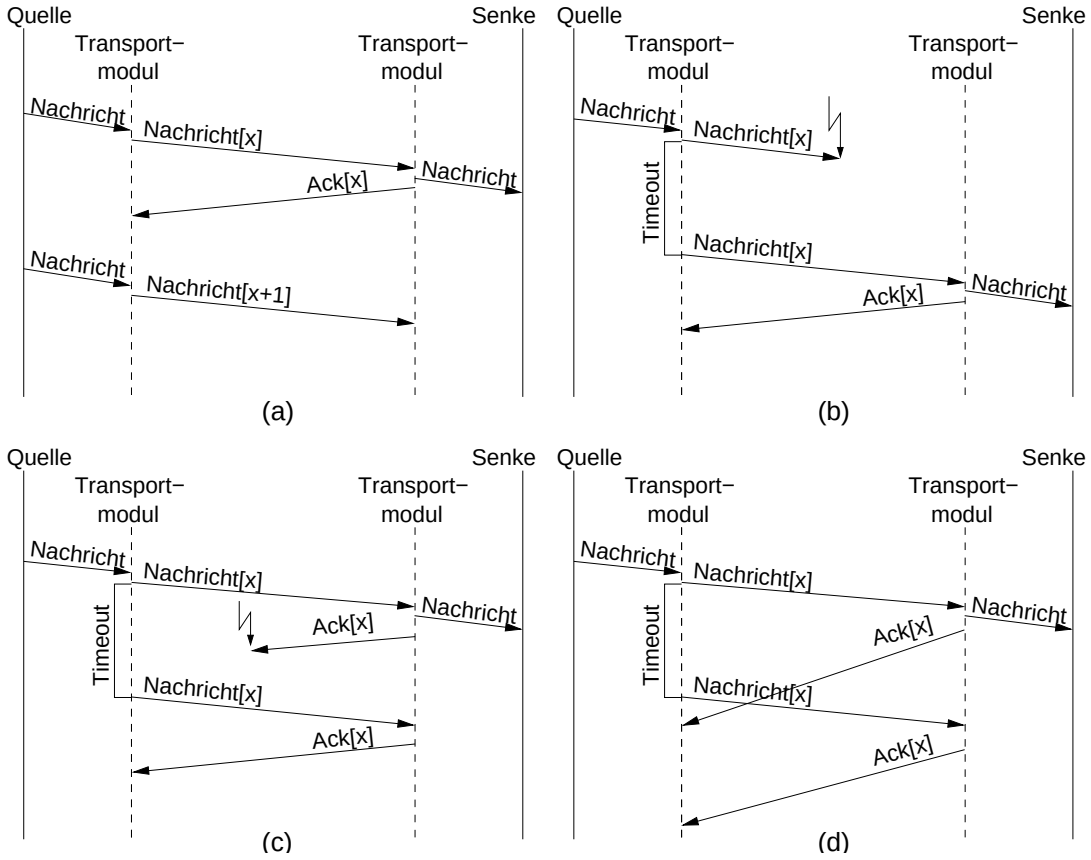


Abbildung 4.4.: Nachrichtentransport: (a) Normalfall. (b) Verlust der Nachricht. (c) Verlust der Bestätigung. (d) Verspätetes Eintreffen der Bestätigung.

Da das Nachrichtenaufkommen bei der Signalisierung eher gering ist und vor allem immer nur einzelne Nachrichten versendet werden, benötigt der Cache wenig Speicherressourcen und es entstehen auch keine Probleme bei überlaufenden Sequenznummern. Aus dem gleichen Grund arbeitet das hier vorgestellte Protokoll ähnlich dem Stop-and-Wait-Protokoll. Das Warten auf eine Bestätigung bevor die nächste Nachricht gesendet werden kann stellt aufgrund des geringen Nachrichtenaufkommens keine Einschränkung dar.

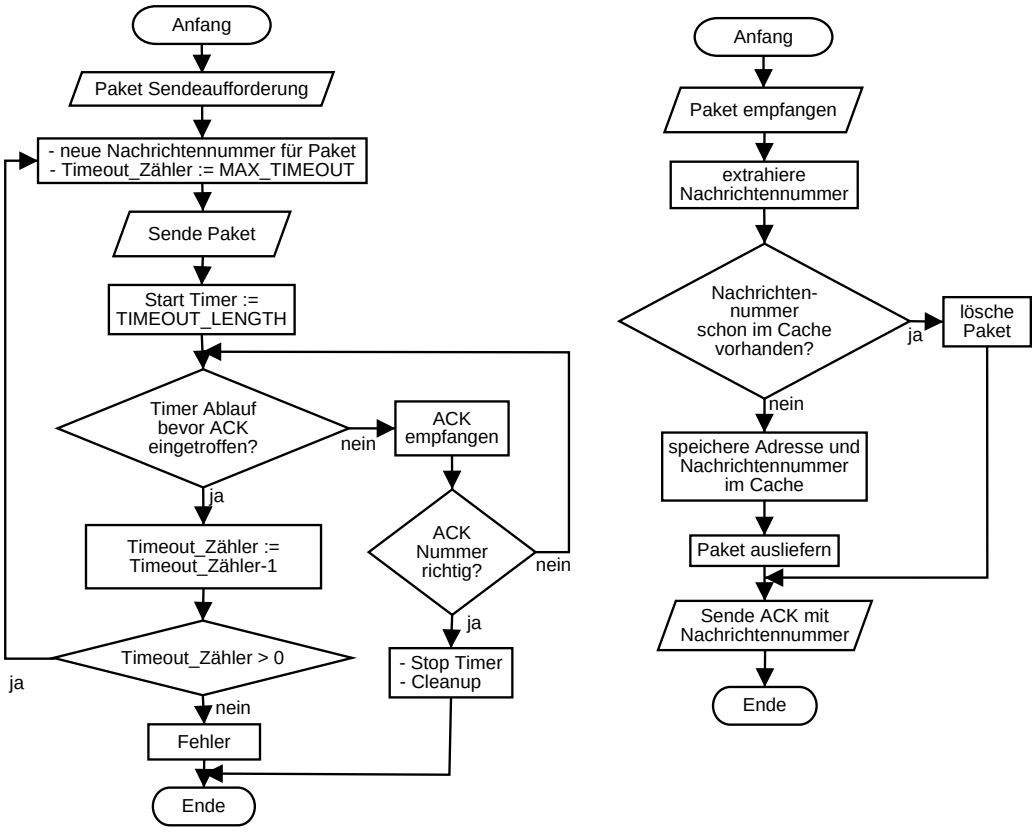


Abbildung 4.5.: Flussdiagramm von Quelle (links) und Senke (rechts).

4.2.3. Verbindungsmanagement

Jede an der Gruppenkommunikation beteiligte Komponente muss in der Lage sein, ATM-Verbindungen aufzubauen und anzunehmen, und das insbesondere auch für mehrere Verbindungen gleichzeitig. Beim Verbindungsaufbau muss beachtet werden, dass dieser Vorgang abgelehnt werden kann. Das kann mehrere Ursachen haben, unter anderem falsche Zieladresse, falscher Service Access Point (SAP), keine ausreichenden Ressourcen für die Verbindung vorhanden (logisch oder physikalisch), etc. Im Unterschied zum Verbindungsaufbau gibt es bei der Annahme von Verbindungen keinerlei Einschränkungen. Das Verbindungsmanagement nimmt alle ankommenden Verbindungen entgegen. Nach Verbindungsannahme wird die neue Verbindung der Anwendung gemeldet und kann dann von dieser akzeptiert oder abgelehnt (also wieder abgebaut) werden.

Alle für die Gruppenkommunikation benötigten Verbindungen werden in zwei Klassen unterteilt: Signalisierung und Datentransport. Signalisierungsverbindungen transportieren ausschließlich Nachrichten, die der Etablierung der Gruppenkommunikation dienen, aber keine eigentlichen Anwendungsdaten. Datentransportverbindungen sind hingegen nur für den Transport von Anwendungsdaten zuständig. Diese Trennung von Signalisierung und Datentransport erlaubt eine organisatorische Trennung der Steuerungslogik der Gruppenkommunikation von dem eigentlichen Datentransport. Diese Trennung

ist dabei nicht nur auf eine Komponente bezogen, sondern erlaubt auch eine physikalische Trennung bzw. Aufteilung von Signalisierung und Datentransport auf mehrere spezialisierte Komponenten.

Ob eine ATM-Verbindung für die Signalisierung oder den Datentransport zuständig ist, wird vom Sender festgelegt, der diese Verbindung aufbaut. Der Sender ordnet der Verbindung einen allgemein bekannten (well known) SAP am Empfänger zu. Ein SAP entspricht im Prinzip einer TCP/UDP-Portnummer und muss im gesamten ATM-Netz eindeutig einem Dienst zugeordnet sein. Um den Missbrauch eines SAP zu verhindern, sollten diese verbindlich festgelegt werden und es sollte auch eine gegenseitige Authentifizierung der Komponenten untereinander erfolgen. Diese Problematik wird in dieser Arbeit weitestgehend außer acht gelassen, da sie nicht speziell mit der Gruppenkommunikation und ATM zusammenhängt.

4.2.4. Zusammenfassung

Auf diesen hier vorgestellten drei Teilen Nachrichtenformat, Nachrichtentransport und Verbindungsmanagement bauen die weiteren Konzepte zur Gruppenkommunikationsunterstützung in ATM-Netzen auf. Durch die Definition eines einheitlichen Nachrichtenformates, das sich aus Basiselementen zusammensetzt, ist eine einfache Erweiterung und Einführung neuer Nachrichten möglich. Alle Komponenten haben dabei das gleiche Modul zur Generierung und Extrahierung von Nachrichten.

Für den Transport der Nachrichten ist ein rudimentäres Transportprotokoll vorgestellt worden. Das Protokoll arbeitet ähnlich dem Stop-and-Wait-Protokoll und es sorgt dafür, dass jede Nachricht nur einmal beim Empfänger ausgeliefert wird.

Das Verbindungsmanagement vereinheitlicht weitere Teile, die in allen Komponenten zur Kommunikation mit ATM benötigt werden. Dieses Modul abstrahiert von der konkreten Angabe spezifischer ATM-Parameter beim Verbindungsaufbau und es übernimmt einen wesentlichen Teil der hierzu nötigen Fehlerkontrollen. Durch die Verwendung dieser drei Module können die im Weiteren vorgestellten Konzepte vereinfacht und auf das Wesentliche konzentriert werden.

4.3. Zusammenfassung

Das Ziel dieses Kapitel ist es, eine Überblick auf die Vorgehensweise bei der Entwicklung von SkaGAN zu geben. Hierzu sind die verschiedenen Teilkonzepte für eine lokale und globale Gruppenkommunikation und deren Zusammenhänge erläutert worden.

Als Grundlage für SkaGAN sind die verwendeten Komponenten (Endsystem, MCS und Controller) im ATM-Netz und ihre Nachrichtenverbindungen beschrieben worden. Für die Kommunikation zwischen den Komponenten ist ein einheitliches und erweiterbares Nachrichtenformat definiert und ein Transportprotokoll für diese Nachrichten eingeführt worden. Darüber hinaus gibt es noch ein einheitliches Modul für die Verwaltung der ATM-Verbindungen, welches ebenfalls von allen Komponenten eingesetzt wird.

5. SkaGAN: Lastverteilung in lokalen ATM-Netzen

In diesem Kapitel wird ein Ansatz [50, 51] beschrieben, der eine Unterstützung für Gruppenkommunikation in lokalen ATM-Netzen ermöglicht [52]. Im Bereich der lokalen ATM-Netze gibt es bereits eine Reihe von Ansätzen (siehe Kapitel 3, ab Seite 29), die sich mit dieser Problematik auseinandersetzen. Alle Ansätze basieren aber entweder auf dem MCS- oder VC-Mesh-Schema oder ermöglichen keine adäquate Lastverteilung (siehe Ansatz in Unterkapitel 3.2.5). In diesem Ansatz wird ein Schema vorgestellt, das eine gute Skalierbarkeit bzgl. der Anzahl der Gruppenteilnehmer in lokalen ATM-Netzen ermöglicht. Hierzu wird das MCS-Schema mit dem Einsatz mehrerer MCS kombiniert. Für die Aufteilung der Gruppenkommunikationsteilnehmer auf die MCS wird ein Lastverteilungsverfahren eingesetzt. Als Gruppenkommunikationsteilnehmer werden im Folgenden verallgemeinert immer Endsysteme bezeichnet, obwohl hierunter auch Multicast-Router verstanden werden können (siehe auch Unterkapitel 4.1, ab Seite 56).

Dieses Kapitel über Gruppenkommunikation in lokalen ATM-Netzen beginnt im folgenden Unterkapitel 5.1 mit einer Definition des Begriffes 'lokales ATM-Netz' und beschreibt darauf zutreffende Netzwerktopologien. Nach welchem Organisationsschema mehrere MCS für die lokalen Gruppenkommunikation eingesetzt werden beschreibt Unterkapitel 5.2. Da der MCS für diesen Ansatz eine zentrale Aufgabe hat, wird dieser in Unterkapitel 5.3 genauer erläutert. Darauf aufbauend wird in Unterkapitel 5.4 das vom Controller verwendete Lastverteilungsverfahren beschrieben. Das letzte Unterkapitel 5.5 enthält eine Leistungsbewertung des hier vorgestellten Ansatzes.

5.1. Definition: lokale ATM-Netze

Von entscheidender Bedeutung für den hier vorgestellten Ansatz ist der Begriff des lokalen ATM-Netzes. Als lokale ATM-Netze werden im Folgenden sternförmige Topologien mit einer zentralen ATM-Schalteinheit angenommen. Die Anzahl der Endsysteme ist bei dieser Topologie durch die Größe (bzw. die Anzahl der Ports) der ATM-Schalteinheit begrenzt. Eine Erweiterung ist die Verbindung mehrerer ATM-Schalteinheiten mittels eines Backbone. Dieser Backbone ist dabei im einfachsten Fall entweder bus- oder ringförmig. So könnten z.B. in einer Firma oder in einem Gebäude jede Abteilung, bzw. jede Etage ein sternförmiges Netz haben und die Abteilungen, bzw. Etagen sind über einen Ring untereinander verbunden (Abbildung 5.1). Mit der Kombination von sternförmigen

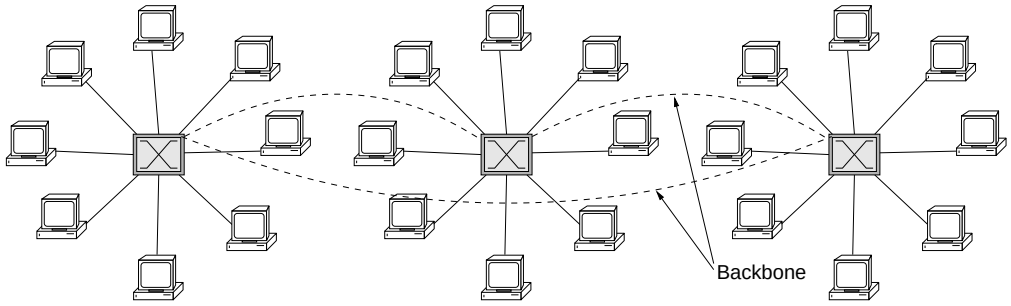


Abbildung 5.1.: Topologie eines lokalen ATM-Netzes.

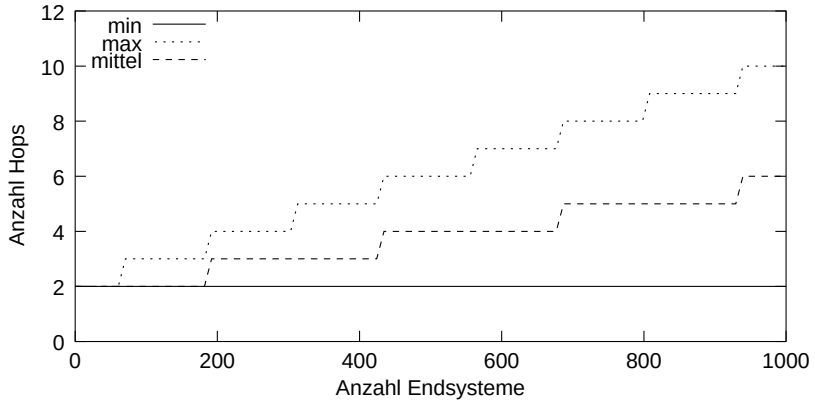


Abbildung 5.2.: Anzahl der Hops zwischen zwei Endsystemen.

Netzen verbunden über eine Backbone-Ringstruktur lassen sich bereits mehrere hundert Endsysteme verbinden, ohne dass es hierbei zu größeren Engpässen kommen muss.

Für die beschriebene lokale ATM-Topologie lässt sich die durchschnittliche Anzahl der Hops zwischen zwei Endsystemen wie folgt bestimmen: Eine ATM-Schalteinheit mit p Ports kann $p-2$ Endsysteme bedienen, zwei Ports werden für den Backbone-Ring benötigt. Bei n Endsystemen sind dann $t = \lceil \frac{n}{p-2} \rceil$ Teilnetze notwendig. Zum x-ten Teilnetz werden $x+1$ Hops benötigt, das eigene Endsystem befindet sich dabei im 1. Teilnetz. Die weiteste Entfernung zweier Teilnetze in einem Ring ist $\lfloor \frac{t}{2} \rfloor$.

In Abbildung 5.2 sind die minimalen, maximalen und mittleren Entfernungen bei Verwendung einer ATM-Schalteinheit mit $p = 64$ Ports dargestellt. Die Treppenstufen ergeben sich an den Stellen, an denen weitere Teilnetze für mehr Endsysteme benötigt werden.

5.2. Kommunikationsschema für mehrere MCS

Bei einem Einsatz mehrerer MCS ist es erforderlich, die Gruppenmitglieder auf die vorhandenen MCS zu verteilen (siehe hierzu auch Kapitel 3.2.5, Seite 39). Bei einem MCS muss dieser alle Sender unterstützen, bei mehreren MCS bieten sich dagegen mehrere Möglichkeiten an, die Gruppenmitglieder auf die MCS aufzuteilen:

Empfänger-orientiert: Eine Möglichkeit besteht darin, dass alle Sender ihre Daten an die MCS weiterleiten, welche Empfänger der Gruppe unterstützen. Jeder MCS bedient nur eine disjunkte Menge von Empfängern dieser Gruppe (Abbildung 5.3(a)).

Jeder MCS erhält weiterhin alle Datenpakete aller Sender, wodurch das Problem der Verkehrskonzentration beim MCS nicht gelöst wird.

Sender-orientiert: Eine andere Möglichkeit ist, die Sender auf die MCS aufzuteilen. Jeder Sender wird dabei genau von einem MCS unterstützt, der die Daten an alle Empfänger einer Gruppe weiterleitet (Abbildung 5.3(b)).

Der Vorteil liegt in der Verminderung des 'Flaschenhalseffektes', da die von den Sendern produzierten Datenmenge auf die vorhandenen MCS aufgeteilt werden kann.

Sender-Empfänger-orientiert: Eine Kombination aus den beiden obigen Ansätzen ist ebenfalls denkbar. Das würde bedeuten, dass ein Sender an alle MCS sendet, und jeder MCS die Datenpakete an alle Empfänger der Gruppe weiterleitet (Abbildung 5.3(c)).

Der Vorteil dieses Ansatzes ist eine hohe Verfügbarkeit der MCS bei Ausfall eines Systems. Es sprechen allerdings wesentliche Nachteile gegen diese Methode. Die Empfänger würden Duplikate erhalten, und zwar so viele, wie MCS vorhanden sind. Diese müssten beim Empfänger wieder entfernt werden und die Duplikate führen zu einer stark erhöhten Netzwerkbelastung. Des Weiteren erhöht sich auch die Anzahl der benötigten ATM-Verbindungen.

Gruppen-orientiert: Eine vierte Alternative wäre eine Einteilung der MCS nach Gruppen. Dabei wird jede Gruppe (alle Sender und Empfänger) genau einem MCS zugeordnet (Abbildung 5.3(d)).

Der Bedarf an ATM-Verbindungen ist hierbei am geringsten (das Datenvolumen bleibt allerdings identisch mit dem Sender-orientierten Ansatz), da jeder Sender und Empfänger nur eine Verbindung zu einem MCS aufbauen muss. Allerdings ist dadurch eine feinere Einteilung nach Gruppenmitgliedern nicht mehr möglich und damit eine Lastverteilung auf die MCS nur unzureichend zu lösen.

Für den Einsatz mehrerer MCS ist der sender-orientierte Ansatz die beste Alternative. Der Bandbreitenbedarf ist am geringsten (wie beim gruppen-orientierten Ansatz) und die Gruppenteilnehmer können so aufgeteilt werden, dass eine verbesserte Lastverteilung auf die MCS möglich wird.

5.3. Modellierung und Bewertung der MCS-Belastung

Dieses Unterkapitel beschreibt ein Warteschlangenmodell für den Bearbeitungsprozess im MCS, mit dem die Datenpaketverzögerung ermittelt werden kann. Hierauf folgt die

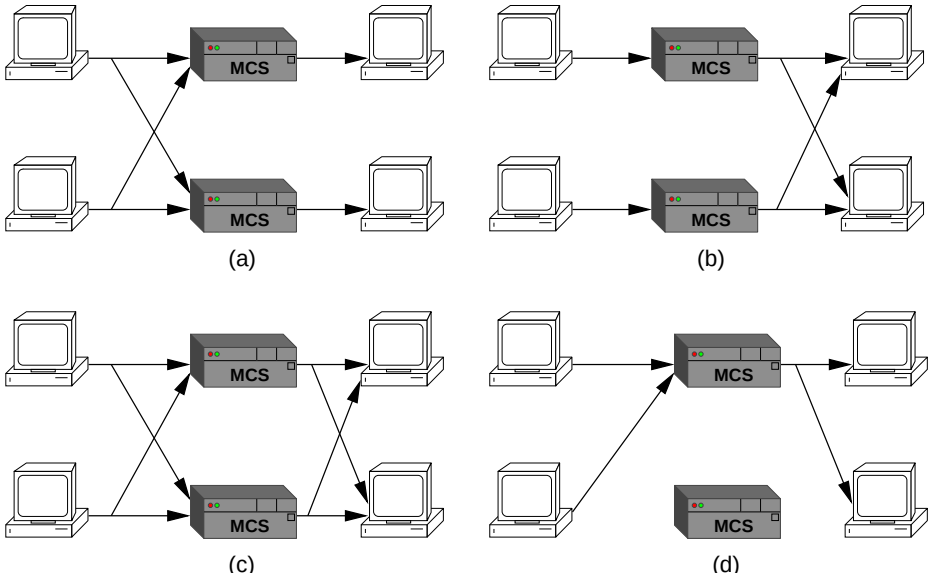


Abbildung 5.3.: Aufteilung der Gruppenteilnehmer auf die MCS: (a) Empfänger-orientiert, (b) Sender-orientierter, (c) Sender-Empfänger-orientiert und (d) Gruppen-orientiert.

Definition eines Maßes für die Belastung eines MCS und eine Leistungsbewertung. Zunächst soll jedoch die interne Verwaltungsstruktur beschrieben werden, da wesentliche Teile des MCS darauf basieren.

5.3.1. Verwaltung

Intern besitzt der MCS eine Datenstruktur wie in Abbildung 5.4 gezeigt. Die Basis bildet eine Liste mit Gruppen, jede Gruppe besteht aus der Gruppenadresse und einer weiteren Liste mit den angemeldeten Sendern der jeweiligen Gruppe. Jeder Sender wiederum kann mehrere Empfängergruppen haben, wobei verschiedene Empfängergruppen auch über verschiedene ATM-Verbindungen versorgt werden.

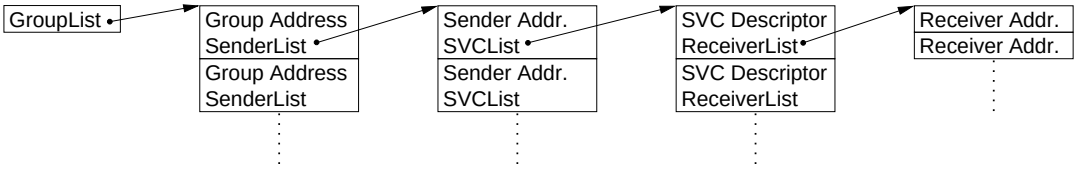


Abbildung 5.4.: Datenstruktur für die Verwaltung im MCS.

Kommt ein Datenpaket von einem Sender bei einem MCS an, so wird zuerst nach der Gruppenadresse in der Gruppenliste des MCS gesucht. Ist diese gefunden, wird die zugehörige Senderadresse bestimmt. Hier kann es mehrere richtige Einträge geben, wenn die Daten an mehrere Empfängergruppen verteilt werden sollen (diese Eigenschaft wird erst für die Datenverteilung bei Weitverkehrsnetzen in Kapitel 6.2 genutzt). Ist der

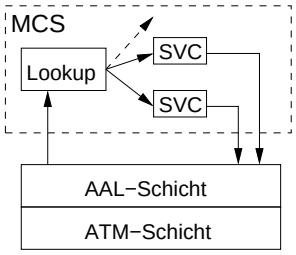


Abbildung 5.5.: Bearbeitung eines Datenpaketes im MCS.

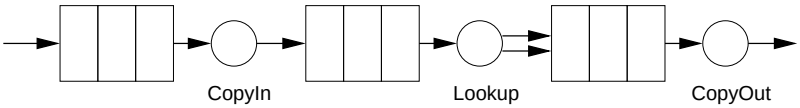


Abbildung 5.6.: Warteschlangenmodell des MCS.

Sendereintrag gefunden, kann das Datenpaket über die zugehörigen ATM-Verbindungen versendet werden.

5.3.2. Warteschlangenmodell

Die Bearbeitung eines Datenpaketes im MCS ist vereinfacht in Abbildung 5.5 dargestellt. Jedes ankommende Datenpaket wird von der ATM-Netzwerkkarte in den Speicher des MCS-Prozesses kopiert. Dann findet ein 'Lookup' statt, die Datenstruktur des MCS wird nach der Gruppe und dem zugehörigen Sender des Datenpaketes durchsucht. Ist der korrekte Eintrag gefunden, und somit auch die zugehörigen ATM-Verbindungen, wird das Datenpaket an die ATM-Netzwerkkarte weitergeleitet.

Um die Bearbeitungsdauer eines Datenpaketes zu bestimmen, wird ein dreistufiges Warteschlangenmodell angenommen, wie in Abbildung 5.6 gezeigt. Die erste Warteschlange (CopyIn) modelliert das Kopieren des Datenpaketes von der ATM-Netzwerkkarte zum MCS-Prozess. Die Bearbeitungsdauer t_{CopyIn} hängt dabei von der Datenpaketgröße ab. Die zweite Warteschlange (Lookup) hat eine konstante Bearbeitungszeit t_{Lookup} pro Datenpaket, um die zugehörigen Einträge in den Datenstrukturen zu ermitteln. Bei einer effizienten Implementierung der Datenstrukturen mittels Hash-Tabellen ist diese Modellierung hinreichend genau, da in einer Hash-Tabelle in annähernd konstanter Zeit ein Eintrag gefunden werden kann. Die letzte Warteschlange (CopyOut) modelliert das Kopieren des Datenpaketes vom MCS-Prozess zur ATM-Netzwerkkarte. Hierfür wird ebenfalls eine von der Datenpaketgröße abhängige Bearbeitungszeit $t_{CopyOut}$ angenommen. Vereinfachend wird weiterhin angenommen, dass die Bearbeitungsdauern für das Kopieren identisch sind: $t_{CopyIn} = t_{CopyOut} = t_{Copy}$.

Zu beachten ist, dass ein Datenpaket mehrfach in der CopyOut-Warteschlange vorkommen kann, wenn es auf mehreren ATM-Verbindungen versendet werden soll. Für die drei Bearbeitungsprozesse werden die folgenden Bearbeitungszeiten angenommen:

Prozess	Dauer
Lookup:	$t_{Lookup} = 100\mu s/Datenpaket$
Copy:	$t_{Copy} = 0,01\mu s/Bit$

Die aus den Bearbeitungszeiten resultierende Verarbeitungskapazität des MCS ist hier sehr konservativ gewählt worden, aktuelle Router und Server haben eine deutlich höhere Verarbeitungskapazität. Die gewählte Verarbeitungskapazität stellt eher eine untere Grenze dar, die von annähernd jedem Endsystem erreicht werden kann.

Für die zu den Bearbeitungsprozessen gehörenden Warteschlangen werden folgende Größen gewählt:

Warteschlange	Größe
Lookup:	$s_{Lookup} = 64 Datenpakete$
CopyIn:	$s_{CopyIn} = 512 KBit$
CopyOut:	$s_{CopyOut} = 512 KBit$

Die Größen der Warteschlangen stellen einen Kompromiss zwischen Paketverzögerung und Paketverlust dar. Die Werte sind so gewählt worden, dass bei einem Verkehrsaufkommen von ca. 80% der MCS-Verarbeitungskapazität die Paketverluste sehr gering sind, aber noch eine akzeptable (geringe) Paketverzögerung vorhanden ist. Nähere Messergebnisse hierzu sind in Unterkapitel 5.5 zu finden.

5.3.3. MCS-Belastung

Im Gegensatz zur Bearbeitungsdauer bzw. Verzögerung eines Datenpaketes im MCS stellt der Faktor Belastung ein Maß für den Arbeitsaufwand eines MCS dar. Das Maß ist äquivalent zum Datenpaketaufkommen an einem MCS und berücksichtigt die drei Bearbeitungsschritte CopyIn, Lookup und CopyOut. Die MCS-Belastung ist vornehmlich ein Maß, um verschiedene MCS miteinander vergleichen zu können. Hierzu ist das Maß für die MCS-Belastung normiert und hat einen Wertebereich von [0...1].

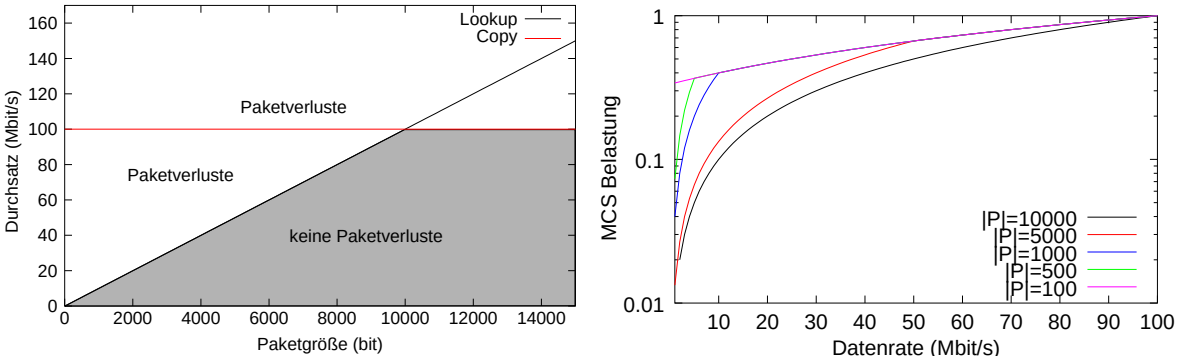
Die Belastung wird wie folgt in regelmäßigen Intervallen von einer Sekunde berechnet:

$$B_{MCS} = \frac{1}{3}(n_{CopyIn} * t_{Copy} + n_{Lookup} * t_{Lookup} + n_{CopyOut} * t_{Copy})$$

Die Variablen sind dabei wie folgt definiert:

B_{MCS}	MCS-Belastung (ohne Einheit)
n_{Lookup}	Anzahl bearbeiteter Datenpakete im Lookup-Prozess (Pakete/s)
n_{CopyIn}	Anzahl kopierter Bits im CopyIn-Prozess (Bit/s)
$n_{CopyOut}$	Anzahl kopierter Bits im CopyOut-Prozess (Bit/s)

Der Maximalwert für die MCS-Belastung ist 1, da die Bearbeitungsprozesse durch die gegebenen Bearbeitungszeiten pro Paket eine Höchstgrenze für die Anzahl der zu bearbeitenden Pakete besitzen. Die Höchstgrenze ist die reziproke Bearbeitungsdauer, der Lookup-Prozess kann also höchstens $n_{lookup}^{max} = 1/t_{lookup}$ Pakete pro Sekunde bearbeiten. Die Anzahlen der bearbeiteten Datenpakete bzw. der kopierten Bits werden im MCS sekündlich gemessen (gezählt) und ausgewertet. Dieser gemessene Wert wird mit



(a) Datenrate und Paketverluste

(b) Belastung

Abbildung 5.7.: Datenraten und Belastungen beim MCS.

vorhergehenden Messungen im Verhältnis 20 zu 80 gewichtet, um Messungenauigkeiten und Schwankungen der Datenraten zu kompensieren:

$$n_x^{neu} = 0,8 * n_x^{alt} + 0,2 * n_x^{mess}; x \in \{CopyIn, Lookup, CopyOut\}$$

5.3.4. Leistungsbewertung

Um den MCS zu bewerten, sind Datenpaketverzögerung, Datenpaketverluste, Datendurchsatz und die errechnete MCS-Belastung von Interesse. Für die folgenden Berechnungen wird vereinfachend angenommen, dass immer nur eine Ausgangsverbindung existiert.

Für die Datenpaketverzögerung im MCS kann ein theoretisches Minimum und Maximum angegeben werden, je nachdem ob die Warteschlangen leer oder voll sind. Da die Bedienprozesse unterschiedliche Bearbeitungskriterien (Datenpakete/s und Bit/s) haben, muss über eine vorgegebene Datenpaketgröße ein Zusammenhang zwischen diesen beiden Kriterien hergestellt werden. Die minimale und maximale Datenpaketverzögerung ergibt sich dann zu (P = Paket, $|P|$ = Paketlänge):

$$\begin{aligned} Delay_{min}(P) &= |P| * t_{Copy} + t_{Lookup} + |P| * t_{Copy} = 100\mu s + 2 * |P| * 0,01\mu s \\ Delay_{max}(P) &= t_{Copy} * s_{CopyIn} + t_{Lookup} * s_{Lookup} + t_{Copy} * s_{CopyOut} + = 11,52ms \end{aligned}$$

Der Datendurchsatz ergibt sich aus der Überlagerung der Werte der drei Bearbeitungsprozesse. Abbildung 5.7(a) zeigt die Durchsatzrate in Abhängigkeit von der Datenpaketgröße. Die Durchsatzrate ist durch den Kopiervorgang auf max. 100 MBit/s begrenzt und durch den Lookup-Bearbeitungsprozess, der maximal 10000 Datenpakete pro Sekunde bearbeiten kann. Die Belastung eines MCS hängt von der Datenrate und der Datenpaketgröße ab, Abbildung 5.7(b) zeigt die Belastungskurven für unterschiedliche Datenpaketgrößen und Datenraten.

5.4. Lastverteilung auf mehrere MCS

Mit Hilfe des Sender-orientierten Kommunikationsschemas (Unterkapitel 5.2) können die Sender unabhängig von einer Gruppe einem MCS zugeordnet werden. Welcher Sender welchem MCS zugeordnet wird, regelt ein Lastverteilungsalgorithmus im Controller. Als zusätzliche Information stehen hierfür noch die Belastungswerte der MCS zur Verfügung. Diese müssen allerdings erst vom Controller bei den MCS abgefragt werden, wozu eine Erweiterung der Signalisierung des MARS Schemas notwendig ist. Zuerst wird in diesem Unterkapitel der verwendete Lastverteilungsalgorithmus motiviert und beschrieben und anschließend die erweiterte Signalisierung.

5.4.1. Lastverteilungsalgorithmus

Die Aufgabe des Lastverteilungsalgorithmus besteht in der Aufteilung der Sender auf mehrere MCS. Das grundsätzliche Problem hierbei ist, dass im Vorhinein nicht bekannt ist, wie lange und mit welcher Datenrate und -charakteristik ein Sender aktiv ist. Daher besteht nur die Möglichkeit, den aktuellen Zustand der MCS zu berücksichtigen, um einen Sender einem MCS zuordnen zu können. Für dieses Zuordnungsproblem sind drei Verfahren näher betrachtet worden, die eine einfache Vergabestrategie ermöglichen:

Round-Robin: Hierbei wird angestrebt, die Sender zu gleichen Anteilen auf die vorhandenen MCS aufzuteilen, so dass alle MCS eine gleich Anzahl von Sendern haben.

Bei diesem Verfahren wird implizit angenommen, dass alle Sender das gleiche Datenaufkommen erzeugen, ansonsten werden die MCS ungleich belastet.

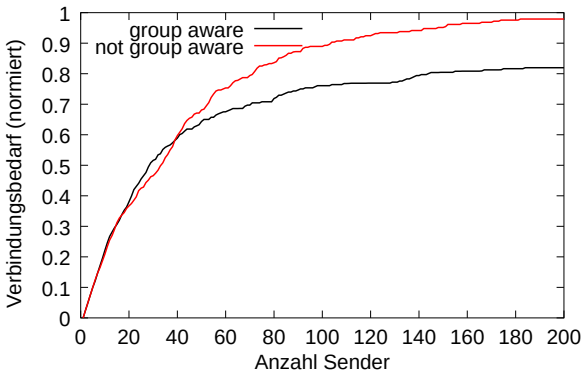
Equal Share: Ein Sender wird dem MCS mit der geringsten Belastung zugeordnet. Hierzu müssen zuerst die Belastungen aller MCS abgefragt werden, um dann denjenigen mit der geringsten Belastung auszuwählen. Die Messung der Belastung in den MCS ist in Unterkapitel 5.3.3 beschrieben.

Dadurch, dass die Last möglichst gleichmäßig auf die MCS verteilt wird, kann auch die zusätzliche Verzögerung, die der MCS verursacht, minimiert werden.

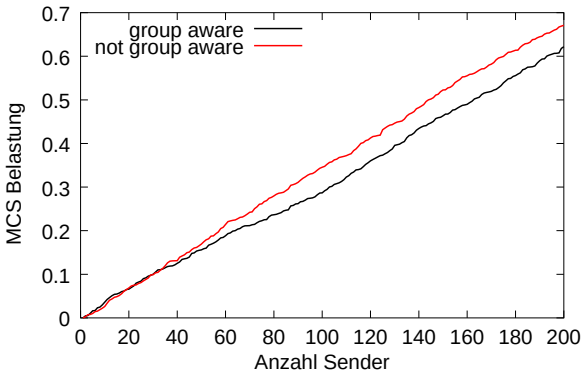
Minimize MCS Usage: Es werden solange Sender einem MCS zugeordnet, bis dieser eine maximale Belastung erreicht hat, erst danach wird zum nächsten MCS gewechselt.

Hierdurch wird die Zahl der benötigten aktiven MCS reduziert, allerdings steigt die Verzögerung in den aktiven MCS, da diese annähernd maximal belastet werden. Des Weiteren ist die Bestimmung der maximalen Belastung eines MCS unklar. Eine absolute Grenze entspräche einem maximal zu verarbeitenden Datenvolumen, womit indirekt auch die Anzahl der Sender begrenzt wäre.

Aus Sicht der Gruppenteilnehmer ist das Equal-Share-Verfahren am besten geeignet, da hierbei die Verzögerung durch die MCS minimiert wird. Das Round-Robin-Verfahren



(a) normierter Bedarf an ATM Verbindungen.



(b) durchschnittliche Belastung der MCS

Abbildung 5.8.: Simulation der Equal-Share-Lastverteilung.

benötigt zu allgemeine Voraussetzungen (alle Sender haben gleiche Datenraten) für eine ausgeglichene Lastverteilung und das Minimize-MCS-Usage-Verfahren erfordert eine maximale Belastungsgrenze der MCS und kann zu hohen Verzögerungen führen.

Unabhängig vom Verfahren ist es aber wichtig festzuhalten, dass die Zuordnung eines Senders zu einem MCS während der gesamten Kommunikationsdauer gleich bleibt. Eine Möglichkeit, einen Sender in diesem Zeitraum einem anderen MCS zuzuordnen, ist nicht vorgesehen. Dies ist nur möglich, wenn sich der Sender zuerst abmeldet und anschließend wieder bei der Gruppe anmeldet. Somit kann prinzipiell nicht vermieden werden, dass, bedingt durch eine Änderung der Senderdatenraten, die Belastung der MCS ungleich verteilt wird. Andererseits wird in so einem Fall ein neu hinzukommender Sender einem weniger belasteten MCS zugeordnet, womit wiederum eine ausgeglichenere Belastung der MCS erreicht wird.

Alle vorgestellten Verfahren haben den Nachteil, dass die Gruppenzugehörigkeit der Sender nicht beachtet wird. Wie im vorherigen Unterkapitel (Gruppen-orientiertes Kommunikationsschema) aber bereits festgestellt wurde, kann die Beachtung der Gruppen zu einer Reduktion der ATM-Verbindungen führen. Um diese Tatsache bei der Senderanmeldung zu berücksichtigen, wird die Belastung eines MCS prozentual geringer gewichtet, wenn die Gruppe des Senders bei dem MCS bereits existiert.

Ein erstes vorläufiges Simulationsergebnis zeigt, dass unter Berücksichtigung der Gruppenzugehörigkeit (Abbildung 5.8) eine durchschnittliche Reduktion der benötigten ATM-Verbindungen von den MCS zu den Empfängern um ca. 20% (Abbildung 5.8(a)) möglich ist, wobei die Belastung der MCS nicht erhöht wird (Abbildung 5.8(b)). Bei der Simulation wurden 10 Gruppen mit jeweils 20 Sendern auf insgesamt 4 MCS zufällig verteilt, die Sender hatten unterschiedliche Sendedatenraten, die ebenfalls zufällig ermittelt worden sind. Die Sender meldeten sich nacheinander bei den Gruppen an und nicht wieder ab. Die Kurven in Abbildung 5.8 stellen die arithmetischen Mittelwerte von je 100 Simulationen dar. Bei den Werten der MCS-Belastung (Abbildung 5.8(b)) sind dabei weniger die absoluten Werte, sondern der annähernd identische Verlauf beider

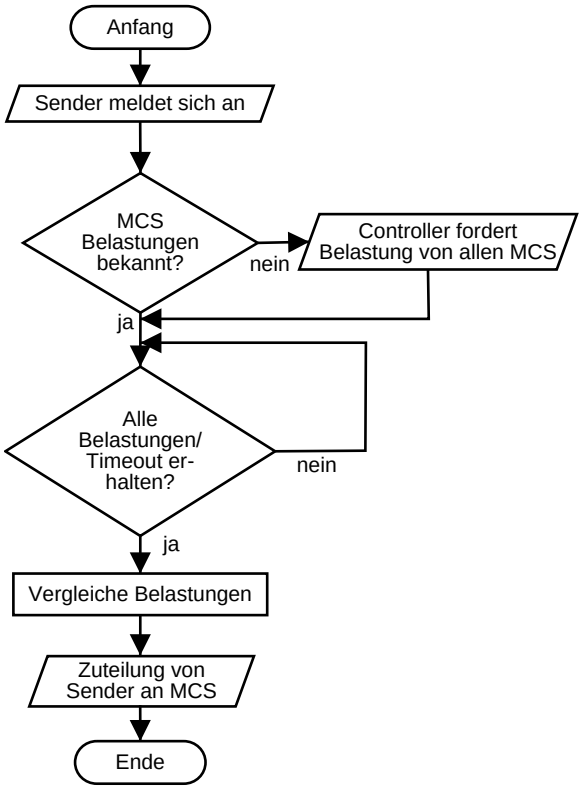


Abbildung 5.9.: Ablaufdiagramm bei Senderanmeldung.

Kurven von Interesse, der bedeutet, dass die Reduzierung der ATM-Verbindungen sich auch positiv auf die MCS-Belastung auswirkt.

5.4.2. Erweiterung der MARS Signalisierung

Der Lastverteilungsalgorithmus wird immer bei einer Senderanmeldung im Controller ausgeführt. Den Ablauf bei einer Anmeldung eines Senders zeigt Abbildung 5.9. Sind die Belastungen der MCS nicht bekannt, werden diese zuerst abgefragt. Anschließend wird der Sender einem MCS zugeordnet. Durch die Abfrage der MCS entsteht eine Verzögerung bei der Anmeldung eines Senders. Diese Verzögerung ist allerdings gering, da zum einen die MCS parallel abgefragt werden und es sich zum anderen um ein lokales ATM-Netz mit kurzen Signallaufzeiten handelt.

Die Abbildung 5.10 gibt den Ablauf der versendeten Nachrichten wieder, die für die Anmeldung eines Senders beim Controller notwendig sind. Nach dem Eintreffen der Anmeldenachricht (Sender_Join) beim Controller verschickt dieser an alle lokal vorhandenen MCS eine Anfrage nach der aktuellen Belastung (Request_Load). Jeder MCS antwortet daraufhin mit seinem aktuell ermittelten Wert (Load_Response). Verloren gegangene Nachrichten werden dabei vom Transportprotokoll aus Abschnitt 4.2.2 behandelt. Haben alle MCS geantwortet, wird derjenige mit der niedrigsten Belastung ausgesucht. Daraufhin wird der Sender und der ausgewählte MCS über den neuen MCS, respektive

Sender informiert.

5.5. Leistungsbewertung

Der in diesem Kapitel vorgestellte Ansatz ist mit dem Netzwerk-Simulator OpNet umgesetzt und modelliert worden. Die drei Komponenten für die Gruppenkommunikation (Endsystem, MCS und Controller) sind in vollem Funktionsumfang (Signalisierung und Datentransport) im Simulator realisiert. Eine Beschreibung von OpNet und den implementierten Komponenten ist in Anhang A ab Seite 163 zu finden.

Es sind zwei Messreihen durchgeführt worden, in denen die wichtigsten Eigenschaften des Gesamtsystems und der Komponenten bewertet worden sind:

1. Das Verhalten des MCS; Paketverzögerung, Paketverluste, Auslastung der Warteschlangen und MCS-Belastungswerte.

2. Simulation des Ansatzes in einem lokalen ATM-Netz. Messung der Lastverteilung und der Ende-zu-Ende-Verzögerung.

5.5.1. MCS

Um sich ein Bild von dem Verhalten des MCS-Modells machen zu können, ist der MCS bezüglich Verzögerung, Datenpaketverlusten, Warteschlangenlängen und Belastung untersucht worden. Als Eingabedatenstrom dient eine einzelne Quelle mit exponentialverteilten Zwischenankunftszeiten und Datenpaketgrößen. Die Exponentialverteilung für die Datenpaketgrößen ist durch die maximal mögliche Paketgröße bei AAL5 nach oben hin begrenzt. Die Exponentialverteilung ist gewählt, da sie unter der Annahme weniger Datenquellen ein hinreichend gutes Modell für die Ankunftszeiten der Pakete liefert [53]. Für eine wachsende Anzahl von Quellen ist eine heavy-tailed-Verteilung besser geeignet, hier sind dann aber auch genauere Annahmen über die beteiligten Datenquellen vonnöten.

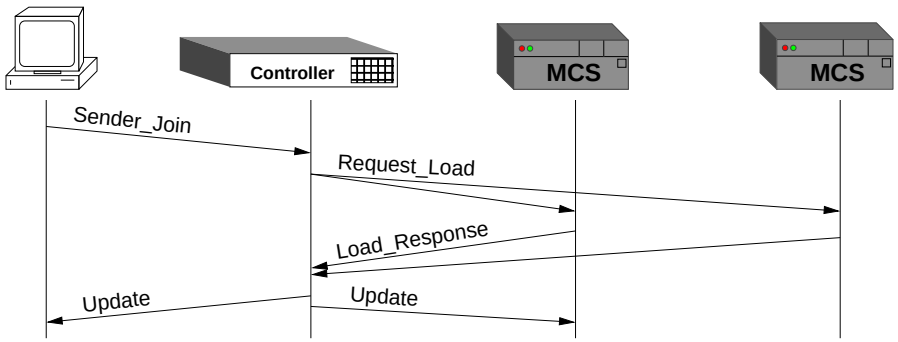


Abbildung 5.10.: Weg-Zeit-Diagramm der Signalisierung bei einem Gruppenbeitritt eines Senders.

Datenrate (MBit/s)	40	50	60	70	80	90	100	110	120	130
Paketgröße (KBit)	7,0	7,5	8,0	8,5	9,0	9,5	10,0	10,5	11,0	11,5
Zwischenankunftszeit(μs)	175	150	133	121	112	105	100	95,4	91,6	88,5
Verzögerung(μs)	391	478	608	813	1247	2416	6387	9615	9773	9624
Warteschlangenlänge										
CopyIn(KBit)	8,16	12,10	18	27,90	46,90	95,70	257	411	458	473
CopyOut(KBit)	5,99	8,08	10,90	14,60	21,90	45,30	144	246	281	307
Lookup(Pakete)	1,00	1,39	1,96	2,81	4,68	9,37	23,50	30,40	23,80	18,10
Paketverluste(MBit/s)	0,00	0,00	0,00	0,00	0,00	0,06	2,02	10,10	19,70	29,20
MCS-Belastung	0,46	0,55	0,65	0,74	0,84	0,92	0,98	0,99	0,99	0,99

Tabelle 5.1.: MCS-Messergebnisse

Mit dem Modell sind für verschiedene Datenraten von 40 bis 130 MBit/s Messwerte ermittelt worden. Die Wahl dieses Wertebereichs ergibt sich aus Abbildung 5.7(a), Seite 71. Ab einer Datenrate über 100 MBit/s und Paketgrößen über 10000 Bit können im MCS theoretisch Paketverluste eintreten. Daher sind die Messreihen um diesen Punkt herum ausgerichtet. Ist die Datenrate deutlich geringer als 100 MBit/s, sind keine Paketverluste zu erwarten. Die Verzögerung ist dann annähernd minimal, da die Warteschlangen kaum gefüllt sind. Bei Datenraten oberhalb 100 MBit/s sind proportional zur Datenrate ansteigende Paketverluste zu erwarten.

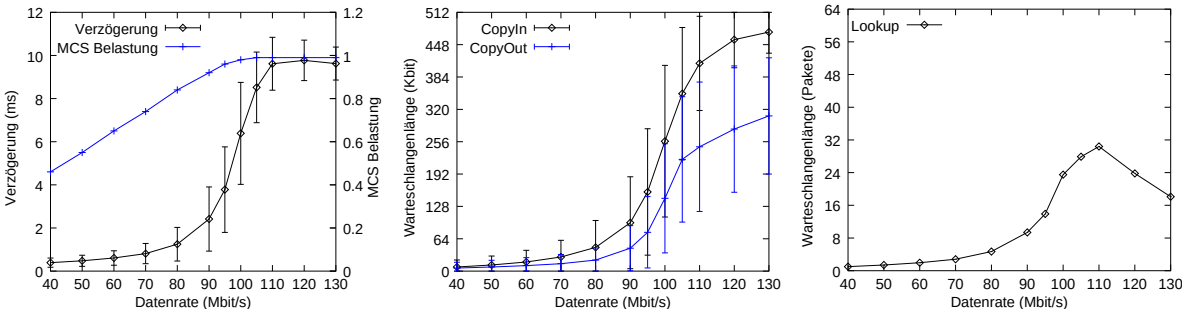
Für jede gewählte Datenrate sind Simulationen von jeweils 8 Sekunden Dauer durchgeführt worden. Aus den Simulationsdaten sind die einzelnen Ergebnisse mit dem arithmetischen Mittel und der Standardabweichung bestimmt worden.

Die Tabelle 5.1 gibt die gemessenen Werte der ersten Messreihe wieder und die Abbildungen 5.11(a) – 5.11(c) stellen die Messwerte mit zusätzlicher Angabe der Standardabweichungen dar. Die ersten drei Zeilen der Tabelle geben die gewählten durchschnittlichen Datenraten, Paketgrößen und Zwischenankunftszeiten wieder, die für die Simulationen gewählt worden sind. Die weiteren Zeilen enthalten die ermittelten Messwerte. Zu beachten ist, dass die (Daten-)Paketverluste in MBit/s angegeben sind, also in der gleichen Maßeinheit wie die Datenrate.

In dieser Messreihe variieren Paketgröße und Zwischenankunftszeit in Abhängigkeit von der Datenrate. In den beiden weiteren Messreihen werden entweder die Paketgrößen (Abbildungen 5.11(d) – 5.11(f)) oder die Zwischenankunftszeiten (Abbildungen 5.11(g) – 5.11(i)) variiert (uniform) und der jeweils andere Parameter konstant gehalten. Zu diesen Messreihen sind keine Wertetabellen angegeben. Die Messwerte in Abbildung 5.11 stellen die arithmetischen Mittel und die Standardabweichungen dar.

Verzögerung

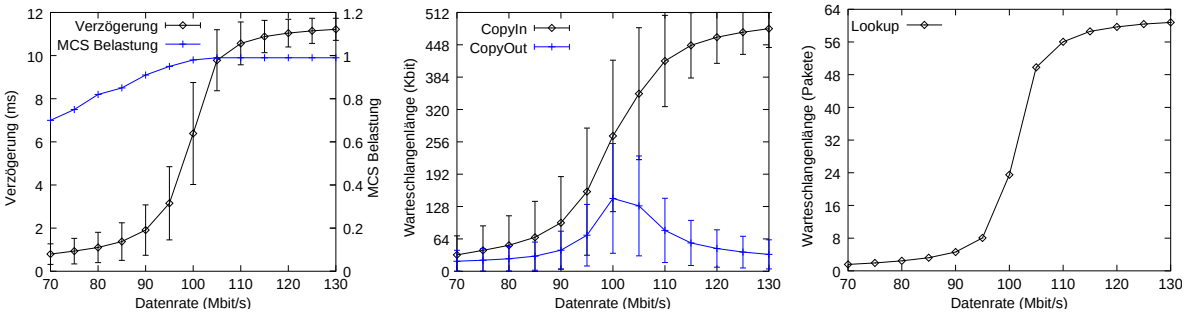
Wie in Abbildung 5.11(a) gut zu erkennen ist, steigt die Verzögerung bis 80 MBit/s Eingangsdatenrate erst sehr langsam an, steigt dann sehr steil, und ist ab Datenraten über 110 MBit/s annähernd konstant. Dies hängt direkt mit dem Ansteigen der Warteschlangenlängen (Abbildungen 5.11(b) und 5.11(c)) zusammen. Die Verzögerung bleibt



(a) Verzögerung und Belastung

(b) Copy-Warteschlangen

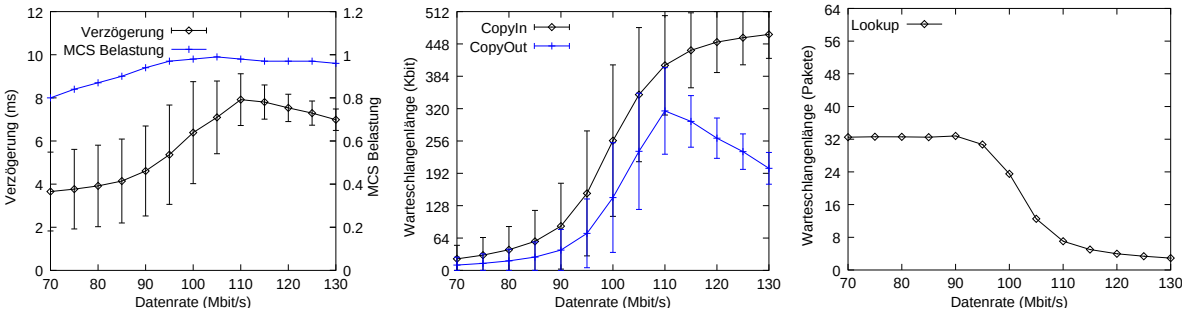
(c) Lookup-Warteschlange



(d) Verzögerung und Belastung

(e) Copy-Warteschlangen

(f) Lookup-Warteschlangen



(g) Verzögerung und Belastung

(h) Copy-Warteschlangen

(i) Lookup-Warteschlangen

Abbildung 5.11.: MCS Simulationsergebnisse: (a), (b) und (c) Variation von Paketgröße und Zwischenankunftszeit, (d), (e) und (f) Variation der Paketgröße (7000-13000bit) bei konstanter Zwischenankunftszeit (=100μs), (g), (h) und (i) Variation der Zwischenankunftszeit (143-77μs) und konstante Paketgröße (10000bit)

oberhalb von 110 MBit/s konstant, da hier die Warteschlangen nahezu vollständig gefüllt sind, und die maximal mögliche Verzögerung ($\simeq 11,52$ ms) erreicht ist. Das ist auch an den angezeigten Standardabweichungen in Abbildung 5.11(a) gut zu erkennen, die bei gefüllten Warteschlangen geringer werden. Die höheren Standardabweichungen im Bereich 80-110 MBit/s geben einen Hinweis darauf, dass die Warteschlangenlängen hier starken Schwankungen unterworfen sind (siehe Abbildung 5.11(b)), und der angegebene Mittelwert keine große Aussagekraft hat.

Warteschlangenlängen

Bei den gemessenen Warteschlangenlängen in Abbildung 5.11(b) ist zu erkennen, dass die CopyIn-Warteschlange immer mehr Daten enthält, als die CopyOut-Warteschlange. Das ist bemerkenswert, da beide Warteschlangen dieselbe Größe und Bearbeitungszeit haben. Der erste Bearbeitungsprozess (CopyIn) funktioniert bei annähernd maximaler Auslastung als eine Art Traffic Shaper, dessen ausgehende Datenrate im Mittel nie größer als 100 MBit/s sein kann. Durch die unterschiedlichen Paketgrößen und den Lookup-Bearbeitungsprozess, der unabhängig von der Paketgröße ist, können nach dem Lookup-Bearbeitungsprozess aber immer noch Bursts auftreten, die einen Überlauf der CopyOut-Warteschlange hervorrufen. Bei Datenraten über 100 MBit/s sinken bei den Warteschlangenlängen die Standardabweichungen, was ebenfalls durch eine Sättigung der Warteschlangen zu erklären ist.

Eine Ausnahme stellt der Lookup-Bearbeitungsprozess dar (Abbildung 5.11(c), dessen Warteschlangenlänge ab 110 MBit/s wieder abnimmt. Es werden im vorherigen CopyIn-Bearbeitungsprozess bei dieser hohen Datenrate so viele Datenpakete verworfen, dass die resultierende Rate der beim Lookup-Bearbeitungsprozess ankommenden Pakete unter 10000 Pakete pro Sekunde fällt und die Warteschlangenlänge vom Lookup-Bearbeitungsprozess kürzer wird. Das ist durch das gleichzeitige Anwachsen von Paketgröße und Verringern der Zwischenankunftszeit bei den Simulationen zu erklären. Dieses Verhalten der Lookup-Warteschlangenlänge ist auch in Abbildung 5.11(i) zu erkennen. Die Zwischenankunftszeiten der Pakete werden verringert, und die CopyIn-Warteschlange füllt sich mit ansteigender Datenrate. Es werden immer mehr Pakete verworfen, wodurch die Warteschlangenlänge des Lookup-Bearbeitungsprozesses kontinuierlich sinkt. Bei einer weiteren Erhöhung der Datenrate sinkt auch die Warteschlangenlänge des CopyOut-Bearbeitungsprozesses (Abbildung 5.11(h)).

MCS-Belastung

Die Belastungskurven des MCS in den Abbildungen 5.11(a), 5.11(d) und 5.11(g) steigen erwartungsgemäß linear mit der Datenrate an. Bei 100 MBit/s ist das Maximum erreicht und die Belastung steigt nicht mehr, obwohl die Datenrate weiter erhöht wird. Die Erklärung hierfür ist, dass verworfene Pakete nicht mit in die Berechnung der Belastung einbezogen werden. Es werden bei der Belastung nur Datenpakete beachtet, die auch von einem der drei Bearbeitungsprozesse bearbeitet werden konnten.

Paketverluste treten bis 90 MBit/s überhaupt nicht oder nur sporadisch auf. Über 100

MBit/s wächst die Rate der Paketverluste im gleichen Verhältnis, wie die ankommende Datenrate über 100 MBit/s liegt. Die resultierende Ausgangsdatenrate des MCS kann das Maximum von 100 MBit/s nicht überschreiten.

Zusammenfassung

Die Belastung und die Verzögerung des MCS hängt von den zu bearbeitenden Paketgrößen und den Zwischenankunftszeiten der Pakete ab. Das hierfür festgelegte Maß der MCS-Belastung spiegelt dieses Verhalten wieder.

Der MCS kann Datenraten bis 80 MBit/s ohne Probleme bearbeiten. Er hat dabei Paketverzögerungen von unter 1,2 ms (siehe Tabelle 5.1) pro Paket und es treten keinerlei Paketverluste auf.

5.5.2. Simulation in einem lokalen ATM-Netz

Um das Verhalten des Ansatzes zur Gruppenkommunikation in lokalen ATM-Netzen beurteilen zu können, werden Messungen zur Lastverteilung auf die MCS und zur Ende-zu-Ende-Verzögerung durchgeführt.

Lastverteilung

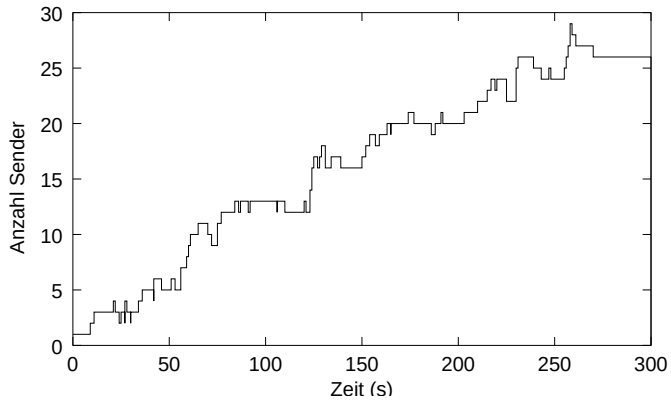
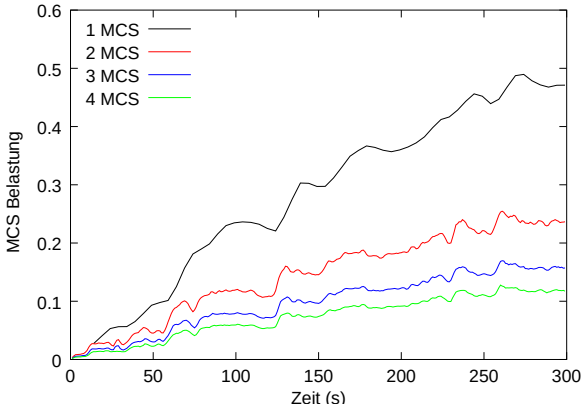


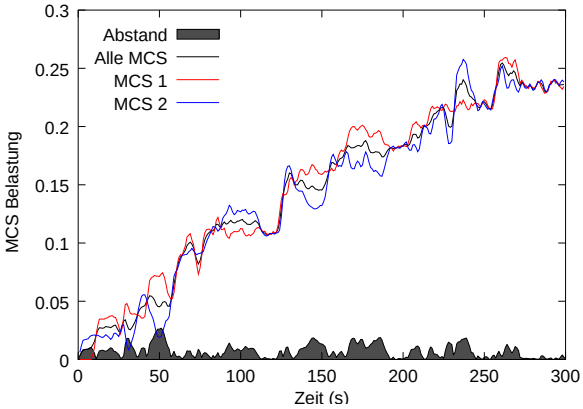
Abbildung 5.12.: Anzahl Sender in der Gruppe während der Simulation.

Um die in Unterkapitel 5.4 vorgestellte Lastverteilung auf mehrere MCS bewerten zu können, ist in erster Linie das Senderverhalten von Bedeutung. Um die Bei- und Austritte von Sendern zu einer Gruppe zu modellieren, wird ein in [54, 55] vorgestelltes Verhaltensmodell zugrunde gelegt. Das Modell basiert auf realen Messungen im MBone und die Teilnehmerbeitritte und Verweilzeiten können mittels zweier verschiedener Verteilungen modelliert werden. Die Abstände zwischen den Teilnehmerbeitritten werden mit einer Exponentialverteilung modelliert und die Verweilzeit eines Teilnehmers wird mit der Zipf-Verteilung [56] angenähert. Da für die Lastverteilung die Sender unabhängig von der Gruppe einem MCS zugeordnet werden, wird nur eine einzelne Gruppe simuliert. Die Gruppe besteht aus 60 Teilnehmern, die in einem Zeitraum von 5 Minuten

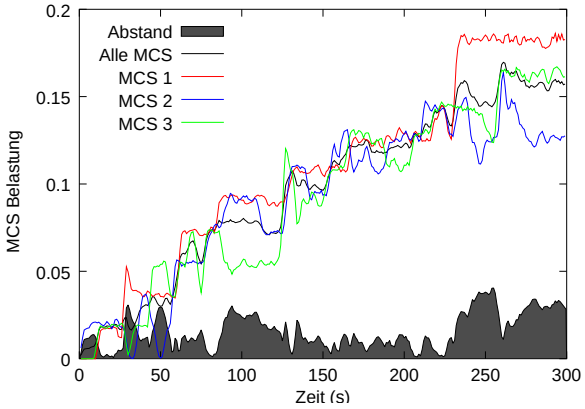
(300 Sekunden) der Gruppe beitreten und zum Teil auch wieder verlassen (Abbildung 5.12). In der Gruppe sind dabei zu einem Zeitpunkt maximal ca. 30 Sender aktiv. Jeder Sender erzeugt eine Datenrate von 5 MBit/s. Die Anzahl der Empfänger einer Gruppe spielt bei der Lastverteilung auf die MCS ebenfalls keine Rolle. Daher existiert für die Simulation nur ein einziger Empfänger in der Gruppe. Die Topologie des ATM-Netzes ist auch ohne Relevanz für die Lastverteilung, da diese bei der Zuordnung der Sender zu einem MCS vom Controller nicht beachtet wird.



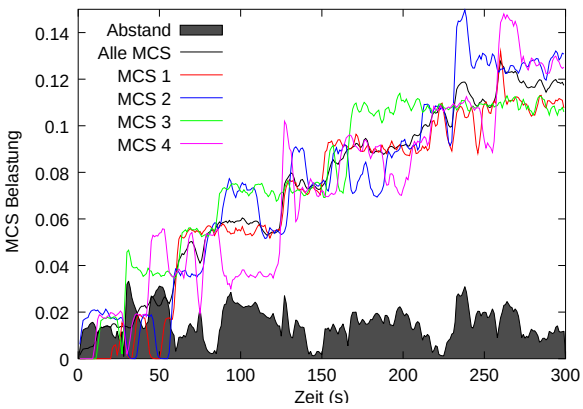
(a) Ein, zwei, drei und vier MCS.



(b) Abweichungen bei zwei MCS.



(c) Abweichungen bei drei MCS.



(d) Abweichungen bei vier MCS.

Abbildung 5.13.: Lastverteilung auf die MCS.

Abbildung 5.13 zeigt die Belastungskurven für das obige Szenario mit ein, zwei, drei und vier aktiven MCS. In der ersten Abbildung 5.13(a) sind zusätzlich zur Belastungskurve bei einem MCS auch die Werte der mittleren MCS-Belastungen aus den Abbildungen 5.13(b)-5.13(d) bei zwei, drei und vier MCS eingezeichnet. Es ist deutlich zu erkennen, dass die Belastung pro MCS im selben Verhältnis sinkt, wie die Anzahl der MCS steigt. Dieser Mittelwert liefert aber keine Aussage darüber, ob die MCS

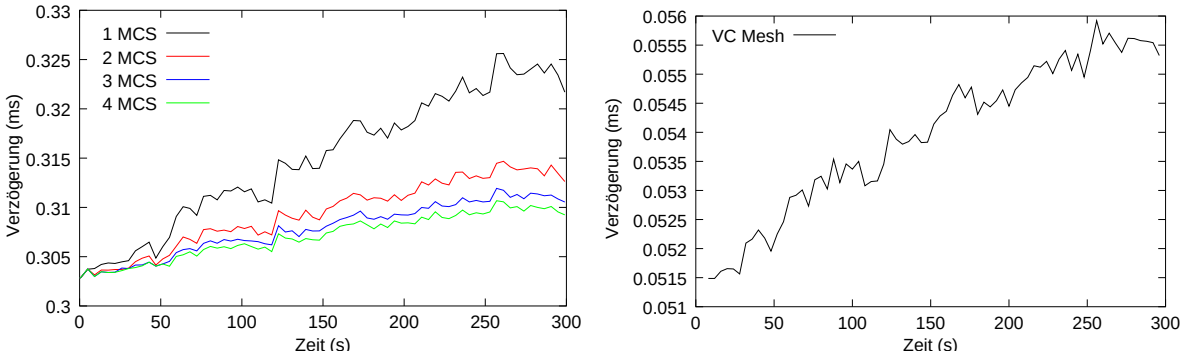
auch gleichmäßig belastet werden, wie es der Lastverteilungsalgorithmus vermuten lässt. Über die Arbeitsweise des Lastverteilungsalgorithmus geben die Abbildungen 5.13(b) - 5.13(d) Auskunft. In den Abbildungen sind jeweils die Belastungswerte der einzelnen MCS wiedergegeben und zusätzlich ist die maximale Abweichung vom Mittelwert als graue Fläche dargestellt. Die Abweichung vom Mittelwert ist als maximale Differenz berechnet und stellt nicht die Standardabweichung dar (da nur wenige Messpunkte pro Zeitpunkt zur Verfügung stehen). Um die Abweichung beurteilen zu können, ist es wichtig zu wissen, dass ein einzelner Sender bei der vorgegebenen Datenrate in einem MCS eine Belastung von ∼0,047 erzeugt. Die dargestellten Abweichungen in den Abbildungen 5.13(b) - 5.13(d) liegen immer deutlich unter diesem Wert. Die Abweichungen der MCS-Belastungen sind also immer geringer als die Belastung, die ein Sender bei einem MCS erzeugt.

Hieraus kann geschlossen werden, dass zum einen die Sender immer auf die MCS mit den geringsten Belastung verteilt werden, und zum anderen dass die Belastung der MCS in gleichem Maße ansteigt wie es der Lastverteilungsalgorithmus vorgibt. Der vorgeschlagene Lastverteilungsalgorithmus ermöglicht also eine ausgeglichene Verteilung der Sender und minimiert somit die entstehenden Verzögerungen und Paketverluste in den MCS.

Ende-zu-Ende-Verzögerung

Aus Sicht eines Gruppenteilnehmers ist die Lastverteilung und die Anzahl der MCS nur indirekt von Bedeutung. Wichtig für die Gruppenteilnehmer ist, wie groß die Verzögerung zwischen Sender und Empfänger ist und welche Verzögerungsschwankungen dabei entstehen können. Es gibt dabei eine Reihe von Faktoren, die Einfluss auf die Verzögerung haben können: Anzahl der MCS, MCS-Belastung, Anzahl Sender, AAL5 Segmentierungsrate und die Paketgröße. Die Leitungsverzögerung kann hingegen vernachlässigt werden, da es sich hier um ein lokales Netz handelt. Die Verzögerung in einer ATM-Schalteinheit wird ebenfalls nicht beachtet und als annähernd konstant angenommen. Erst bei einer sehr hohen Netzwerkbelastung entsteht eine erhöhte Verzögerung in einer ATM-Schalteinheit, die aber bei den hier vorliegenden Messungen nicht auftritt.

Die Messung der Ende-zu-Ende-Verzögerung mit dem gleichen Szenario wie bei der Lastverteilung zeigt Abbildung 5.14. Wie nicht anders zu erwarten, steigt die Verzögerung mit der Anzahl der Sender und sinkt mit der Anzahl der MCS (Abbildung 5.14(a)). Es gilt also: Verzögerung $\sim \alpha \frac{AnzahlSender}{AnzahlMCS}$ mit α als Skalierungsfaktor. Die minimale Verzögerung beim Einsatz eines MCS ist hier $\sim 300\mu$s. Dieser Wert ergibt sich aus der Bearbeitungszeit des MCS, der Paketgröße und der Segmentierungsrate. Daraus errechnet sich die minimale Verzögerung zu: $2 * \frac{5000bit}{135MBit/s} + 100\mu s + 2 * 5000 * 0,01\mu s = 277\mu s$. Die Differenz aus dem gemessenen und dem berechneten Wert ist $\sim 23\mu s$ und ist die Verzögerung auf der Leitung und in der ATM-Schalteinheit. Zum Vergleich zeigt Abbildung 5.14(b) die Verzögerung beim VC-Mesh-Schema, bei dem die Daten direkt von den Sendern zu den Empfängern gesendet werden. Die Verzögerung steigt hier ebenfalls mit der Anzahl der Sender an, ist aber im Vergleich erheblich niedriger, da die Verzögerung im MCS entfällt und die Datenpakete nur einmal beim Sender segmentiert werden



(a) Verzögerung mit MCS.

(b) Verzögerung mit VC Mesh.

Abbildung 5.14.: Ende-zu-Ende-Verzögerung bei veränderlicher Senderanzahl.

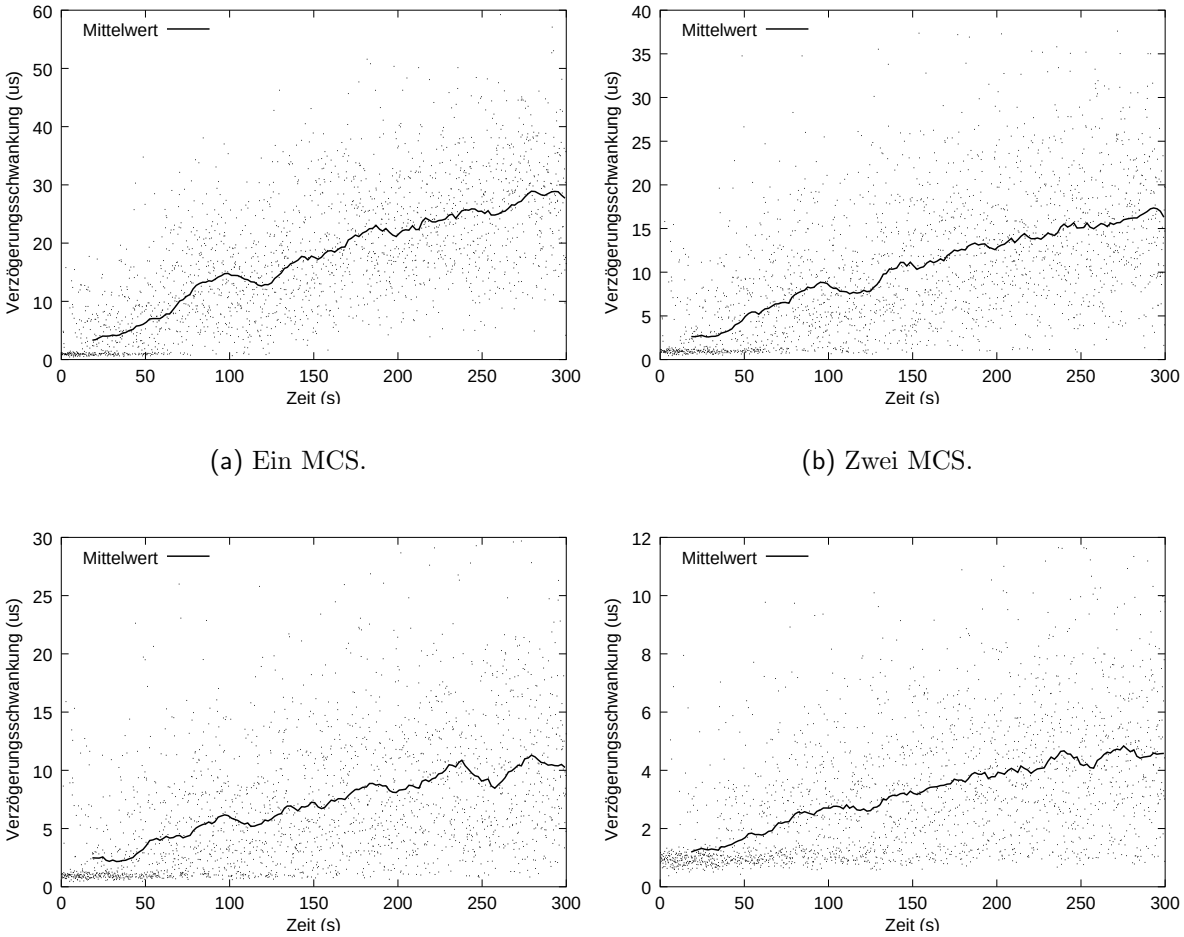
müssen.

Dem Vorteil der geringeren Verzögerung des VC-Mesh-Schemas steht allerdings der erhöhte Signalisierungsaufwand bei Gruppenänderungen entgegen (siehe auch Unterkapitel 2.4, Seite 23). Der erhöhte Signalisierungsaufwand setzt sich aus der MARS Signalisierung und der ATM-UNI Signalisierung zusammen. Dabei nimmt die UNI Signalisierung den größeren Teil ein, jeder Sender muss seine ATM-Verbindung zu den Empfängern der Gruppe ändern.

Ergänzend zur Verzögerung gibt Abbildung 5.15 die Verzögerungsschwankungen wieder. Die Diagramme in Abbildung 5.15 haben unterschiedliche Skalen an der y-Achse, um die Wertebereiche besser darstellen zu können. Analog zur Verzögerung steigen die Verzögerungsschwankungen ebenfalls mit der Senderanzahl und somit mit der Datenrate an. Je mehr MCS aktiv sind, desto geringer werden die Schwankungen. Die geringsten Schwankungen treten beim VC-Mesh-Schema auf. Gegenüber der Verzögerung in Abbildung 5.14(a), die linear mit der Anzahl der MCS gesunken ist, nehmen die Verzögerungsschwankungen nicht linear mit der Anzahl der MCS ab. In den Diagrammen 5.15(a) - 5.15(c) liegen die Verzögerungsschwankungen mit zwei MCS bei 60% gegenüber denen bei einem MCS und bei 40% bei vier MCS. Darüber hinaus ist zu bemerken, dass bedingt durch die AAL5 Segmentierung und die Segmentierungsgeschwindigkeit immer geringe Verzögerungsschwankungen entstehen, die zu einem Versatz auf der y-Achse der Diagramme in Abbildung 5.15 führen.

5.6. Zusammenfassung

In diesem Kapitel ist der Einsatz mehrerer MCS in lokalen ATM-Netzen beschrieben worden. Um hieraus einen größtmöglichen Nutzen ziehen zu können, wurde ein Lastverteilungsalgorithmus vorgestellt, der die Gruppenteilnehmer dynamisch den MCS zuordnet. Damit wird die Belastung gleichmäßig auf die MCS verteilt, was aus Sicht der Empfänger in den Gruppen zu reduzierten Verzögerungen, Verzögerungsschwankungen



(a) Ein MCS.

(b) Zwei MCS.

(c) Vier MCS.

(d) VC Mesh.

Abbildung 5.15.: Verzögerungsschwankungen bei unterschiedlichen MCS Anzahlen.

und Paketverlusten führt.

Um dieses Systemverhalten untersuchen und belegen zu können, ist zunächst ein Warteschlangenmodell des MCS erstellt und ein normiertes Maß für dessen Belastung definiert worden. Durch ein erweitertes MARS-Signalisierungsprotokoll werden jetzt zuerst die Belastungen der MCS ermittelt, bevor ein Sender einem der MCS zugeordnet wird. Damit trotz des Einsatzes mehrerer MCS weiterhin alle Daten korrekt ausgeliefert werden, ist ein Kommunikationsschema entworfen worden, dass keine zusätzliche Netzwerkbelastung hervorruft, bzw. dieselbe Belastung wie das MCS-Schema mit einem MCS erzeugt.

6. SkaGAN: Gruppenkommunikation in ATM-Weitverkehrsnetzen

Für eine Gruppenkommunikationsunterstützung in Weitverkehrsnetzen [57, 58] sind eine Reihe von Anforderungen zu erfüllen, die in lokalen Netzen (Kapitel 5) nur bedingt relevant sind. Hierzu zählt vor allem die Skalierbarkeit bezüglich der Gruppengröße und der Gruppentopologie. Aber auch die Dynamik einer Gruppe und Aspekte der Ausfallsicherheit spielen eine Rolle bei Weitverkehrsnetzen.

Während in lokalen Netzen die Topologie einer Gruppe im Allgemeinen durch die Netzwerktopologie vorgegeben und annähernd konstant ist, bestehen bei Weitverkehrsnetzen keinerlei direkte Zusammenhänge zwischen Gruppen- und Netzwerktopologie mehr. Je nach Gruppengröße partizipieren nur Teilbereiche des Gesamtnetzwerkes an der Kommunikation innerhalb einer Gruppe. Hieraus folgt auch, dass verschiedene Gruppen unterschiedliche Netzsegmente nutzen können, was bei lokalen Gruppen nahezu ausgeschlossen ist. Dieser Aspekt ist wichtig, da er es ermöglicht, Datenkonzentrationen in Komponenten bei Weitverkehrsnetzen zu vermeiden. Durch die Dynamik einer Gruppe, also die Änderung der Teilnehmer einer Gruppe, kann sich ebenfalls die Gruppentopologie im Laufe der Zeit ändern. Das erfordert eine kontinuierliche Anpassung der Kommunikationswege, da ansonsten der Datentransfer das Netzwerk unnötig belasten würde.

Bei der Ausfallsicherheit ist es wünschenswert, dass nur die direkt vom Netz- oder Serverausfall betroffenen Komponenten bei der Gruppenkommunikation unterbrochen werden. Bei einem lokalen Netz bedeutet das, dass z. B. bei dem Ausfall eines MCS alle an diesen angeschlossenen Teilnehmer (Sender) keine Daten mehr übertragen können, alle anderen Sender aber vom Ausfall nicht betroffen sind. Diese lokale Begrenzung beim Ausfall einer Komponente soll auch für die Gruppenkommunikation in Weitverkehrsnetzen gelten. Als Teilnehmer werden hier, wie beim lokalen Ansatz (Kapitel 5), Endsystem oder Multicast-Router verstanden (siehe auch Unterkapitel 4.1, ab Seite 56).

Als Basis für die Gruppenkommunikation wird in Unterkapitel 6.1 zuerst ein Schema für die Verwaltung von Gruppen in Weitverkehrsnetzen beschrieben, das die oben beschriebenen Anforderungen erfüllt. Auf dieser Verwaltung basiert der Datentransfer zwischen den Mitgliedern einer Gruppe, der in Unterkapitel 6.2 dargestellt wird. Unterkapitel 6.3 stellt zwei Erweiterungen vor, die eine bessere Anpassung an dynamische Gruppenänderungen ermöglichen und für große Gruppen die Lastverteilung optimieren können. Unterkapitel 6.4 schließt das Kapitel mit einer Leistungsbewertung ab.

6.1. Gruppenverwaltung

Eine Voraussetzung für die Gruppenkommunikation ist eine Verwaltung und Lokation der Gruppen im ATM-Netzwerk. Als Erstes ist es wichtig, eine Gruppe identifizieren und adressieren zu können. Hierzu werden in diesem Ansatz Gruppenadressen zur Adressierung einer Gruppe eingesetzt, denn diese Form der Adressierung hat durch IP-Multicast die bisher weiteste Verbreitung und Akzeptanz gefunden. Die Gruppenadresse dient zur Identifikation der Gruppe, darüber hinaus ist es aber notwendig, die Teilnehmer einer Gruppe zu verwalten und zu organisieren, so dass gewährleistet ist, dass an die Gruppe gesendete Daten auch in jedem Fall alle zugehörigen Teilnehmer erreichen können. Es wird hierzu eine Abbildung im Netzwerk benötigt: $Gruppenadresse \rightarrow Teilnehmeradressen$.

Von der Umsetzung dieser Abbildung hängt die Effektivität der Gruppenverwaltung ab. Insbesondere sollten bei der Umsetzung die folgenden Faktoren beachtet werden:

Sichtbarkeit: Die geografische Begrenzung der Sichtbarkeit von Gruppenadressen, auch als Scoping bezeichnet. Das erlaubt die Mehrfachverwendung von Gruppenadressen und ermöglicht vor allem die Beschränkung des Datentransfers auf ein Teilnetz.

Lebensdauer: Gruppenadressen sind gewöhnlich nicht permanenter Natur, wie z. B. Unicast-Adressen (es gibt auch permanente Adressen, diese haben dann eine unendliche Lebensdauer). Daher ist die Definition der Lebensdauer einer Gruppenadresse wichtig. Die Begrenzung der Lebensdauer einer Gruppenadresse ist vor allem für eine spätere Wiederverwendung der Adresse von entscheidender Bedeutung.

Skalierbarkeit: Der Verwaltungsaufwand im Verhältnis zur Gruppengröße. Die Verwaltung soll sowohl für kleine als auch für große Gruppen effizient arbeiten. Der Nachrichten- und Verarbeitungsaufwand soll dabei höchstens linear mit der Gruppengröße ansteigen.

Der hier realisierte Ansatz zur Gruppenverwaltung ist an das PNNI-Routing[20] angelehnt. Beim PNNI-Routing werden Netzsegmente zusammengefasst und als ein abstraktes Netzsegment auf der nächst höheren Ebene dargestellt. Diese Hierarchiebildung wird ebenfalls für die Gruppenverwaltung eingesetzt. Mehrere lokale Netze werden zu einem abstrakten Segment zusammengefasst. Für jedes dieser Segmente muss für die Verwaltung ein Controller vorhanden sein. Ein Beispiel zeigt Abbildung 6.1(a). Die Endsysteme sind die Blätter des Baums und die Knoten sind die Controller, die für das Gruppenmanagement verantwortlich sind. Dargestellt ist ein Baum mit zwei Ebenen, weitere Ebenen sind nach dem gleichen Grundprinzip zu bilden. Die mögliche Realisierung der hierarchischen Gruppenverwaltung in einem ATM-Netzwerk zeigt die Abbildung 6.1(b). Wie in der Abbildung zu sehen ist, muss ein Controller nicht eindeutig einer ATM-Schalteinheit zugeordnet sein. Ein Controller kann ein ATM-Endsystem sein oder auch in einer ATM-Schalteinheit integriert werden. Diese Flexibilität wird erreicht, indem sämtliche Kommunikation zwischen dem Controller und der ATM-Schicht über die ATM-UNI-Schnittstelle durchgeführt wird.

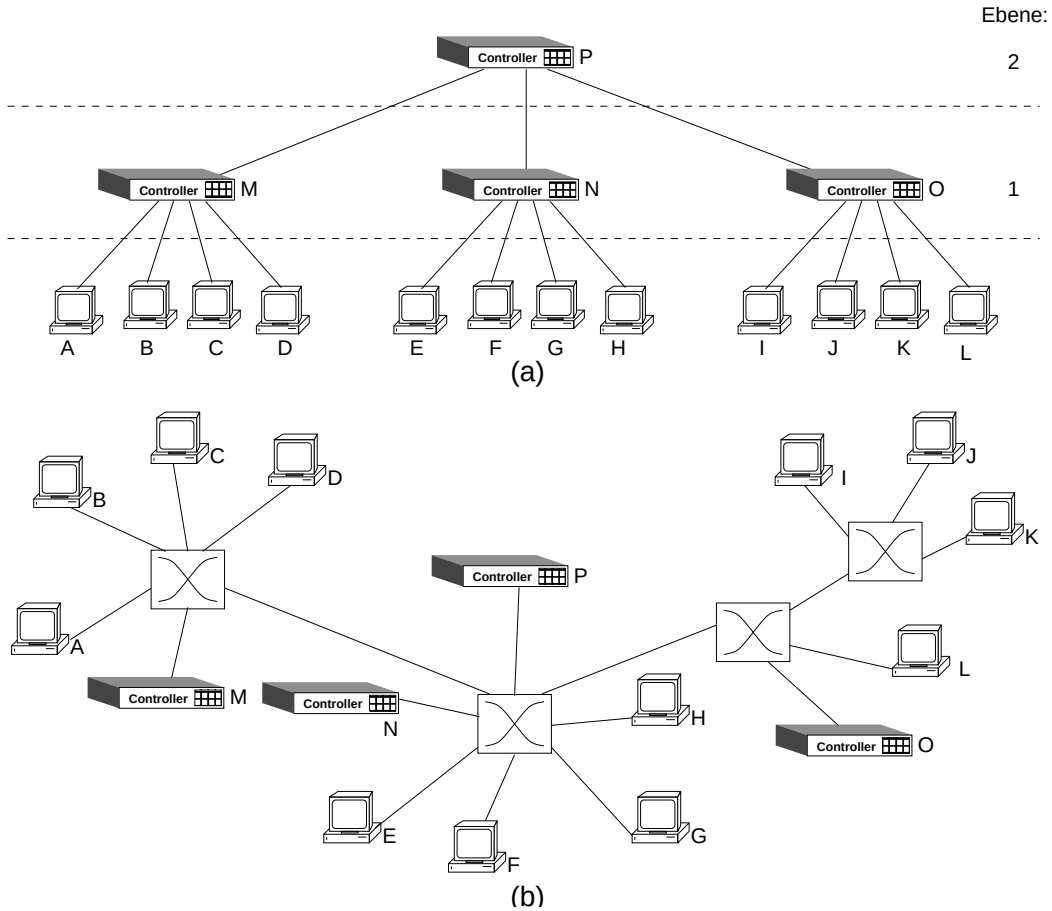


Abbildung 6.1.: Hierarchische Gruppenverwaltung: (a) Baumstruktur und (b) Realisierung der Baumstruktur.

Die Hierarchieebenen orientieren sich an der Definition der Sichtbarkeitsbereiche von ATM-Anycast-Adressen, wie sie vom ATM-Forum für UNI 4.0 [21] festgelegt sind. Diese Sichtbarkeitsbereiche stellen ein Äquivalent zu den IP Multicast Address Scopes [4] dar. Durch die zusätzliche Angabe der Sichtbarkeit kann eine Gruppenadresse auf eine vorgegebene Hierarchieebene beschränkt werden.

Die Controller haben die Aufgabe, die Gruppenadressen auf eine Menge von ATM-Adressen abzubilden und vor allem die Informationen über die Gruppen bzw. Teilnehmer zu sammeln, sie zu aggregieren und mit anderen Controllern auszutauschen. Der Austausch der Informationen erfolgt immer nur zwischen Controller und Endsystem des gleichen lokalen Netzes oder zwischen Controllern benachbarter Hierarchieebenen.

Für den Informationsaustausch zwischen benachbarten Controllern werden ATM-Verbindungen (SVCs) etabliert, welche den notwendigen Signalisierungsverkehr transportieren. Für die Etablierung der SVCs sind die Adressen der Controller notwendig, die vorher konfiguriert werden müssen. Zur Erleichterung der Konfiguration werden ATM-Anycast-Adressen verwendet. D. h. zur Etablierung eines SVCs wird die Hierarchieebene sowie die ATM-Anycast-Adresse benötigt.

Im Vergleich zu den Aufgaben beim lokalen Ansatz von SkaGAN (Kapitel 5) ist jetzt im Wesentlichen die hierarchische Gliederung zwischen den Controllern und der damit verbundene Intormationsaustausch hinzugekommen. Die Kommunikation zwischen Endsystem und Controller und die Abbildung der Gruppenadresse auf eine Menge von Teilnehmeradressen ist hingegen gleich geblieben.

6.1.1. Etablierung einer Verwaltungshierarchie

Die Hierarchie für die Gruppenverwaltung sollte sich an der Topologie des Netzes orientieren, ist aber im Prinzip nicht daran gebunden, da es nur eine logische Struktur darstellt, die annähernd beliebig auf ein konkretes Netz abgebildet werden kann. Die Verwaltungshierarchie ist dabei permanent vorhanden und stellt eine Art von Gerüst oder Overlay-Netz dar, über das Informationen zwischen den beteiligten Gruppenkommunikationskomponenten ausgetauscht werden können.

Die einzelnen Verwaltungsknoten (Controller) im Baum sind über SVCs verbunden. Hier stellt sich das Problem, dass entweder jeder Knoten die Adresse des Vaterknotens im Baum oder umgekehrt jeder Knoten die Adressen aller seiner Söhne kennen muss. Die Kenntnis der expliziten Adressen erfordert aber einen hohen Konfigurationsaufwand und ist außerdem nicht fehlertolerant. Bei Ausfall eines Knotens müsste wieder per manueller Konfiguration ein anderer Knoten ausgewählt werden.

Eine Lösung hierfür bietet ATM-Anycast, eingeführt bei UNI 4.0. Anycast-Adressen dienen dazu, einen Dienst im Netzwerk anzusprechen ohne genaue Kenntnis über die Lokation des Dienstes. Die Anycast-Adressen für bestimmte Dienste müssen daher vorher vereinbart und im ATM-Netz eindeutig festgelegt sein. Mehrere Knoten können sich für eine Anycast-Adresse anmelden. Beim Verbindungsaufbau wird versucht, immer den nächstgelegenen Knoten als Verbindungsziel zu nehmen. ATM-Anycast-Adressen haben darüber hinaus noch eine Reichweite (Address Scope), die die Gültigkeit dieser Adressen auf einen Netzbereich beschränkt.

Für jede Hierarchieebene der Gruppenverwaltung wird je eine ATM-Anycast-Adresse reserviert. Das sind insgesamt 15 Adressen, eine für jeden vom ATM Forum [21] festgelegten Address Scope. Diese Adressen sind allen Controllern bekannt, die an der Gruppenkommunikation teilnehmen. Es ist jetzt nur noch notwendig, den einzelnen Knoten entsprechende Hierarchieebenen zuzuordnen, und somit implizit eine ATM-Anycast-Adresse. Außerdem muss in den Knoten den jeweiligen Anycast-Adressen noch ein Address Scope zugeordnet werden, um die Sichtbarkeit zu begrenzen. Dieser Scope ist von der Netzwerktopologie abhängig und muss daher in jedem Knoten entsprechend angepasst werden. Die jeweiligen Hierarchieebenen der einzelnen Controller müssen manuell konfiguriert werden. Ein Signalisierungsprotokoll, dass den Baum für die Gruppenverwaltung etabliert, existiert nicht. Die manuelle Konfiguration jedes einzelnen Controllers ist auf den ersten Blick nicht sehr praktikabel. Aber diese Konfiguration muss nur einmalig für eine gegebene Netzwerktopologie durchgeführt werden. Des Weiteren ist eine manuelle Konfiguration notwendig, um der Hierarchieebene des Controllers und der damit implizit verbundenen Anycast-Adresse einen Address Scope zuzuordnen. Dieser Address Scope ist immer von der Netztopologie abhängig. Da über die Topologie keine

Informationen in den Komponenten existieren, ist eine automatische Konfiguration in den Komponenten nur mit einem großen Aufwand zu realisieren. Eine manuelle Konfiguration ist darüber hinaus sehr einfach durchzuführen, wenn die PNNI-Hierarchie beachtet wird.

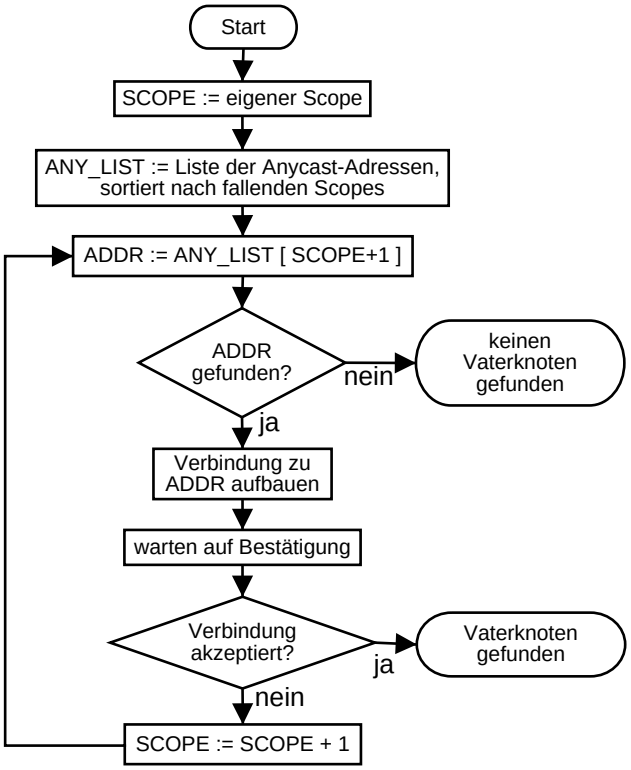


Abbildung 6.2.: Flussdiagramm für den Aufbau einer Verbindung zu einem Vaterknoten.

Die Baumhierarchie entsteht von unten nach oben, indem Sohnknoten SVCs zu Vaterknoten aufbauen. Hierzu wird in jedem Knoten der in Abbildung 6.2 gezeigte Algorithmus ausgeführt. Der Algorithmus wird mit dem eigenen Scope (und damit implizit auch der eigenen Anycast-Adresse) und der absteigend sortierten Liste aller Anycast-Adressen für die Hierarchieebenen initialisiert. Danach wird versucht eine Verbindung zum nächst höheren Baumknoten aufzubauen. Dieser Vorgang wird wiederholt, bis entweder eine Verbindung zu einem höheren Knoten etabliert werden konnte oder kein höherer Knoten gefunden worden ist. Beim letzteren Fall ist dann der Knoten zwangsläufig der Wurzel-Knoten des Baumes.

Eine Voraussetzung für das Verfahren ist, dass die Vaterknoten ihre jeweiligen Anycast-Adressen registriert haben müssen, bevor ein Sohnknoten eine Verbindung aufbauen kann. Insbesondere muss ein gemeinsamer Wurzelknoten existieren, ansonsten entsteht eine Menge von unverbundenen Teilbäumen. Diese Bedingung muss bei der manuellen Konfiguration der Controller beachtet werden, damit die Gruppenverwaltung im gesamten ATM-Netz funktionieren kann.

Registriert sich ein Knoten im Baum verspätet, entsteht eine Situation, wie in Ab-

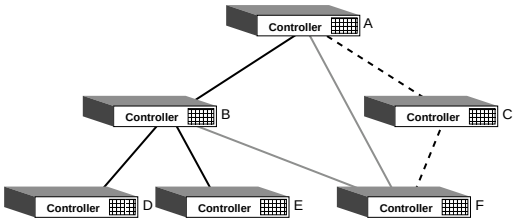


Abbildung 6.3.: Baumaufbau bei unvollständiger Registrierung der Anycast-Adressen.

bildung 6.3 dargestellt. Der Knoten C meldet sich verspätet an. Das führt dazu, dass der darunter befindliche Knoten F sich entweder am nächst höher gelegenen Knoten A registriert, oder bei B anmeldet, wenn der Scope der Anycast-Adresse von Knoten B weit genug reicht.

Dieses Fehlverhalten, welches zu einem Baum führen kann, der bei der Konfiguration der einzelnen Knoten nicht erwünscht war, kann korrigiert (optimiert) werden, indem der Aufbau-Algorithmus (Abbildung 6.2) in regelmäßigen Abständen (z. B. stündlich) ausgeführt wird. Wurde dabei ein neuer, direkterer Knoten im Baum gefunden, so wird die bisherige Verbindung zum alten Vaterknoten abgebaut und der neu gefundene Knoten ist dann der Vaterknoten.

Ausfall eines Controllers

Das hier vorgestellte Verfahren kann ohne Änderung auch bei Ausfall eines Controllers angewendet werden. Dabei sind zwei Fälle zu unterscheiden:

1. Der Controller ist ein Blattknoten. In dem Fall ist nur der lokale Controller, aber keine weiteren Knoten in der Hierarchie betroffen. Der Controller wird beim Vaterknoten abgemeldet.

2. Der Controller ist Knoten im Baum. Er hat somit einen Vater- und mehrere Sohnknoten. Der Vaterknoten meldet den Controller ab und die Sohnknoten des Controllers löschen diesen ebenfalls aus Ihrem Datenbestand. Jeder Sohnknoten führt daraufhin den in Abbildung 6.2 gezeigten Aufbau-Algorithmus aus, um einen neuen Vaterknoten zu finden.

Der Ausfall eines Controllers wird dem Vater- und den Sohnknoten über die ATM-Signalisierung mitgeteilt (siehe Unterkapitel 4.2.3, Seite 63), da die ATM-Verbindung zwischen den beiden Knoten nicht mehr aufrecht erhalten werden kann.

Ein Sonderfall ist noch der Ausfall des Wurzelknotens. Dieser hat keinen Vaterknoten, was dazu führt, dass die Sohnknoten der Wurzel getrennte Teilbäume bilden. Eine Lösung hierfür ist die Einführung eines Backup-Knotens, der in der Hierarchie über dem Wurzel-Knoten angeordnet ist, wie in Abbildung 6.4 gezeigt.

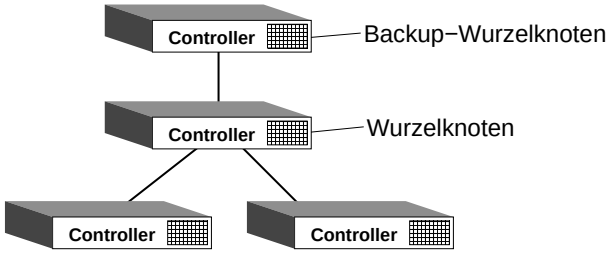


Abbildung 6.4.: Der Ausfall des Wurzelknotens kann durch einen Backup-Wurzelknoten kompensiert werden.

6.1.2. Signalisierung und Verwaltung im Controller

Der Aufbau einer hierarchischen Struktur bildet für SkaGAN die Basis für die Gruppenverwaltung, denn über den Verwaltungsbaum sind alle Controller untereinander verbunden. Alle die Gruppenverwaltung betreffenden Nachrichten werden über diese Hierarchie weitergeleitet. Aber initiiert werden Änderungen bei einer Gruppe immer von einem Endsystem. Ein Endsystem kann einer Gruppe lokal als Sender und/oder Empfänger beitreten oder diese Gruppe verlassen. Hierfür können die in Kapitel 5 vorgestellten Nachrichtenpakete und Prozeduren für die An- und Abmeldung der Teilnehmer eingesetzt werden. Für die Weiterleitung einer Gruppenänderung im Verwaltungsbaum sind allerdings weitere Mechanismen notwendig.

Bei der Verwaltung der Teilnehmer wird explizit zwischen Sendern und Empfängern einer Gruppe unterschieden. Diese Unterscheidung hat für die Verwaltung selbst keinerlei Vorteile, die Informationen sind aber für den Datentransfer (Unterkapitel 6.2) von Bedeutung. Das Grundprinzip bei der Weiterleitung ist, dass nur Änderungen (Neueintrag, Löschung) einer Gruppe zur nächst höheren Hierarchieebene weitergegeben werden.

Dieses Verhalten ist in Abbildung 6.5 dargestellt. Die gezeigten Endsysteme und Controller beziehen sich dabei auf Abbildung 6.1, Seite 87. Das Weg-Zeit-Diagramm stellt die Signalisierung für eine einzelne Gruppe dar, daher sind die sonst notwendigen Gruppenadressen zur Vereinfachung weggelassen. Für eine Gruppe stellt das Diagramm aber alle wesentlichen Fälle dar, die bei der Signalisierung eintreten können. Der Gruppe treten nacheinander die Endsysteme A, B und E als Empfänger bei (RJ = Receiver Join). Anschließend kommt Endsystem F als Sender hinzu (SJ = Sender Join) und der Empfänger E verlässt am Ende die Gruppe wieder (RL = Receiver Leave).

Das Endsystem A tritt der Gruppe bei, indem es eine ReceiverJoin-Nachricht (das Format ist in Anhang C ab Seite 177 beschrieben) an den lokalen Controller M sendet. Der Controller M erklärt sich daraufhin ebenfalls zum Empfänger für dieselbe Gruppe und meldet das dem nächst höheren Controller P, der dann auch Empfänger für diese Gruppe wird. Die nächste Anmeldung von Empfänger B hat nur noch lokale Auswirkung beim Controller M, denn dieser ist bereits als Empfänger für die Gruppe registriert. Der Empfängerbeitritt von Endsystem E läuft analog zum Beitritt von Endsystem A ab. Bei dem Beitritt von Endsystem F zur Gruppe ist nur der Status als Sender im Controller N geändert. Jetzt ist auch Controller N Sender in der Gruppe, was dieser dem Controller P

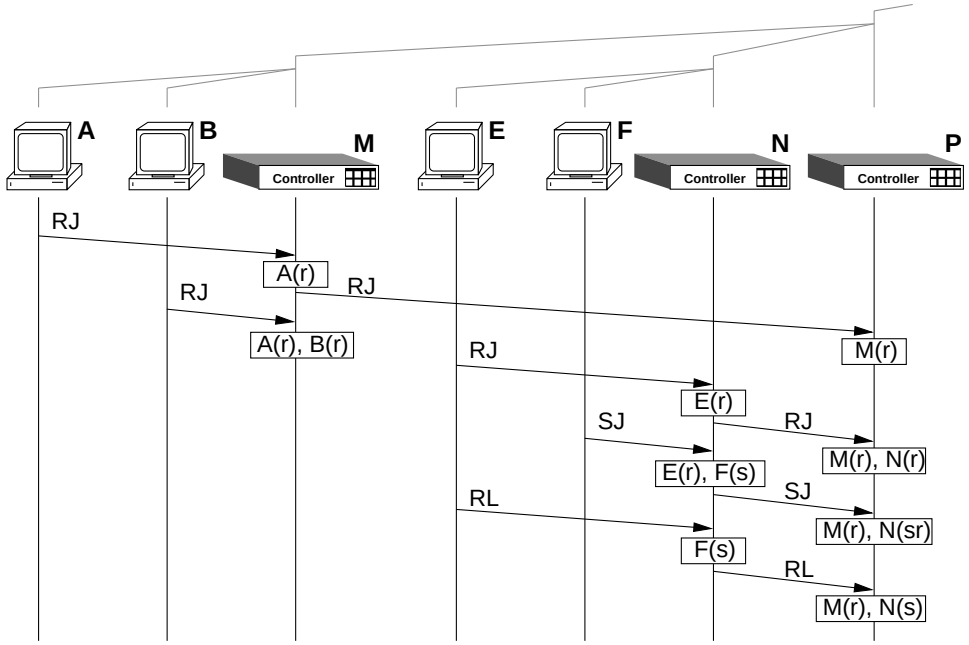


Abbildung 6.5.: Weg-Zeit-Diagramm mit Signalisierung bei Gruppenänderungen.

signalisiert. Im Controller P erhält System N den Status als Empfänger und Sender. Das Verlassen der Gruppe von Endsystem E bewirkt auch eine Abmeldung vom Controller N bei Controller P als Empfänger.

Bewertung

Der Aufwand für die Speicherung der Gruppendaten in den Controllern und die Anzahl der Signalisierungsnachrichten kann durch die hierarchische Gruppenverwaltung signifikant verringert werden. Andererseits hat jeder Knoten im Baum immer nur eine Teilsicht auf eine Gruppe. Jeder Knoten besitzt nur die Information, welcher seiner Teilbäume Sender oder Empfänger für die Gruppe enthält, die genauen Adressen sind nicht bekannt. Das führt zu einer stark verteilten Datenhaltung, bei der jeder Knoten eine andere Sicht auf die Teilnehmer einer Gruppe hat. Andererseits wird durch diese Informationsreduktion auch eine gute Skalierbarkeit erreicht.

Bei dem Ausfall eines Controllers entsteht bei dem vorgestellten Schema nur ein temporärer Informationsverlust. Die gespeicherten Informationen im Controller sind immer nur Abstraktionen, die aus den Informationen der darunter liegenden Controller gewonnen werden. Der ausgefallene Controller wird im darüberliegenden Controller aus der Gruppe entfernt, was im schlechtesten Fall noch weitere Änderungen in höheren Hierarchieebenen nach sich ziehen kann. Die darunter liegenden Controller versuchen Verbindungen zu anderen Controllern aufzubauen und ihre jeweiligen Gruppen bei den neuen Controllern erneut zu registrieren.

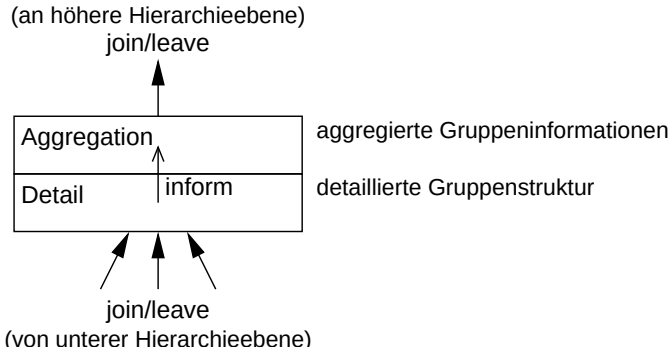


Abbildung 6.6.: Zwei-Schichten-Modell des Controllers.

Funktionales Modell des Controllers

Für die Verarbeitung der Nachrichten zur Gruppenverwaltung ist der Controller in zwei Schichten unterteilt, wie in Abbildung 6.6 dargestellt. Die untere Detailschicht speichert alle Daten über die Gruppenzugehörigkeiten der unteren Teilbäume. Bei Eintreffen einer Sender/Receiver-Join/Leave-Nachricht werden die Datenbestände aktualisiert. Wird dabei eine Gruppe neu eröffnet oder gelöscht, so wird die höhere Aggregationsschicht informiert. In der Aggregationsschicht wird nur gespeichert, in welchen Gruppen der Controller als Sender oder Empfänger vertreten ist. Wenn die Aggregationsschicht von der darunter liegenden Detailschicht über eine Gruppenänderung informiert wird, wird die zugehörige Gruppe aktualisiert und eine Nachricht über die Gruppenänderung an den nächst höheren Controller gesendet.

In dieses Schema kann sehr einfach die Berücksichtigung der Sichtbarkeit von Gruppenadressen integriert werden. Nur wenn die Sichtbarkeit einer Gruppenadresse größer ist als die Hierarchieebene des Controllers, wird die Aggregationsschicht von der Detailschicht informiert. Ist die Sichtbarkeit der Gruppenadresse mit der Hierarchiehöhe identisch, so wird die Gruppe nur in der Detailschicht aktualisiert und ist in höheren Schichten nicht mehr sichtbar.

6.2. Datentransfer

Im vorhergehenden Unterkapitel ist die Gruppenverwaltung beschrieben worden, die Informationen über Gruppen sammelt, speichert und über eine Baumstruktur austauscht. Diese Informationen werden benötigt, um darauf den Datentransfer zwischen Gruppenteilnehmern etablieren zu können. Für den Datentransfer wird ein Schema verwendet, das festlegt, wie die ATM-Verbindungen zwischen den Gruppenteilnehmern aufgebaut werden müssen. Den eigentlichen Transport der Nachrichten im Netz übernimmt dann ATM. Dieses Schema wird in Unterkapitel 6.2.1 beschrieben. Danach wird das Signalisierungsprotokoll in Unterkapitel 6.2.2 vorgestellt, das, basierend auf der Gruppenverwaltung, eine verteilte Organisation ermöglicht. Die einzelnen Aktionen und Tätigkeiten im Controller, die für den reibungslosen Ablauf des Signalisierungsprotokolls nötig sind,

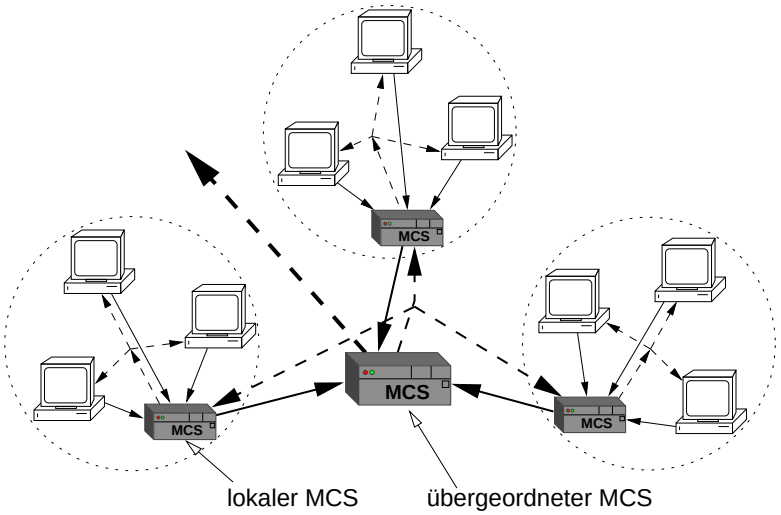


Abbildung 6.7.: Hierarchisches MCS-Schema.

beschreibt Unterkapitel 6.2.3.

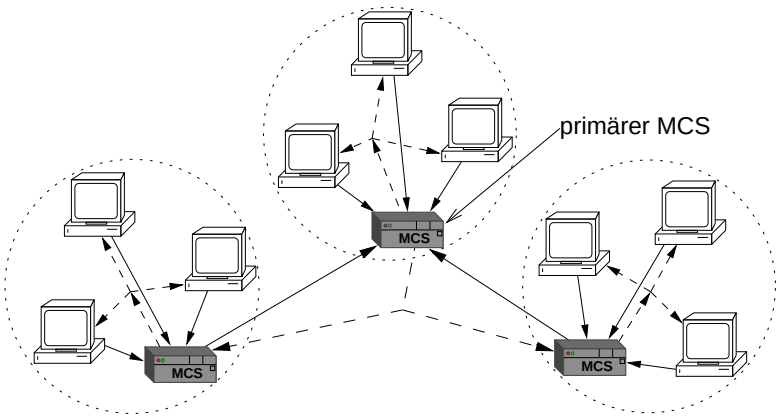
6.2.1. Gruppenkommunikationsschema

Die Einführung einer Hierarchie ist eine naheliegende Möglichkeit zur skalierbaren Gruppenkommunikationsunterstützung in Weitverkehrsnetzen, unabhängig von ATM. Die bei der Gruppenverwaltung eingeführte Baumstruktur ist dafür allerdings ungeeignet. Für die Gruppenverwaltung existiert nur ein einziger statischer Baum, welchen sich alle Gruppen, und somit alle Sender, teilen müssten. Das würde im Fall des Datentransfers zu einer Datenkonzentration in den Knoten führen, insbesondere in den höheren Hierarchieebenen.

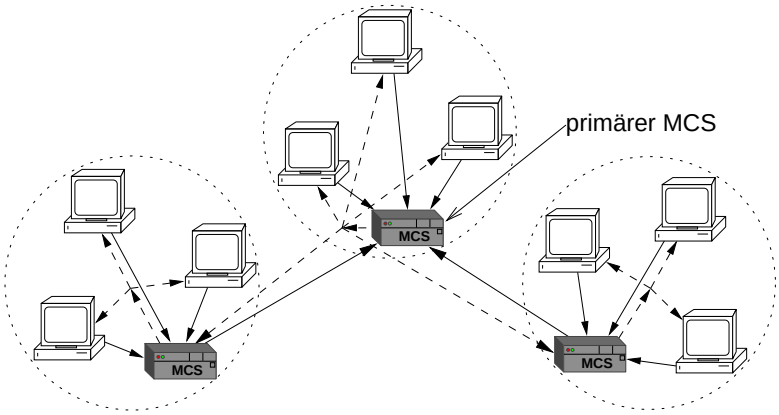
Ein anderer Nachteil bei der Gruppenverwaltung ist, dass die Anzahl der Zwischensysteme unnötig hoch sein kann. Ein Zwischensystem ist hier ein MCS, in dem die Datenpakete in der Anwendungsschicht bearbeitet werden. Das erhöht die Ende-zu-Ende-Verzögerung, da jedes Mal die Datenpakete in der AAL-Schicht segmentiert und reassembliert werden müssen damit sie von der Anwendungsschicht bearbeitet werden können. Durch die im Vorhinein festgelegte Baumstruktur ist die Anzahl der Zwischensysteme zwischen zwei Gruppenteilnehmern bestimmt, unabhängig von der Gruppengröße. D. h. nur die geografische Verteilung der Gruppenteilnehmer entscheidet über die notwendige Anzahl der Zwischensysteme, nicht die Gruppengröße. Für eine effiziente Unterstützung der Gruppenkommunikation wäre es hingegen sinnvoll, die Anzahl der Zwischensysteme möglichst gering zu halten, bzw. nur in Relation zur Gruppengröße ansteigen zu lassen.

Ein ähnliches Schema wie bei der Gruppenverwaltung kann aber dennoch verwendet werden. Abbildung 6.7 zeigt die Methodik der Hierarchiebildung auf das MCS-Schema angewandt. Für jede Hierarchieebene existiert ein dedizierter MCS. Dieser MCS verteilt die Daten der Teilnehmer in den jeweiligen Hierarchieebenen. Dieser erste Ansatz

verlangt den zusätzlichen Einsatz von MCS und die Hauptnachteile des MCS, der Flaschenhalseffekt bei vielen Sendern und der MCS als Single Point of Failure (vgl. Kapitel 3), verstärken sich bei diesem Schema. Ein weiterer Nachteil ist im Fall von vielen aktiven Gruppen vorhanden. Während ein lokaler MCS nur die Gruppen bedient, in denen seine Teilnehmer aktiv sind, muss ein MCS in höheren Hierarchieebenen alle Gruppen bedienen, die in den Unterbäumen aktiv sind. Das führt mit ansteigender Hierarchieebene zu einer Erhöhung der Gruppenanzahl und des Datenvolumens in den MCS.



(a) Nutzung eines vorhandenen MCS für die nächste Hierarchieebene.



(b) Wiederverwendung der lokalen ATM-Verbindungen für den Datentransfer zu anderen MCS.

Abbildung 6.8.: Verteiltes hierarchisches MCS-Schema.

Das Problem kann wesentlich reduziert werden, wenn ein lokaler MCS die Aufgaben eines übergeordneten MCS mit übernimmt, wie in Abbildung 6.8(a) dargestellt. Dieser MCS, der für die lokalen Teilnehmer und die höheren Hierarchieebenen gleichzeitig zuständig ist, wird im Weiteren auch als *primärer MCS* bezeichnet (für die verschiedenen MCS Bezeichnungen siehe auch Anhang E). Die Datenkonzentration durch viele

aktive Gruppen in einem MCS kann jetzt vermieden werden, indem ein primärer MCS pro Gruppe gewählt wird. Somit können unterschiedliche Gruppen auf mehrere primäre MCS verteilt werden.

Eine weitere Optimierungsmöglichkeit existiert in der Reduktion der ausgehenden ATM-Verbindungen eines primären MCS (siehe Abbildung 6.8(b)). Es wird nur noch eine ausgehende ATM-Verbindung beim primären MCS benötigt, über die die Gruppendaten sowohl an die lokalen Teilnehmer als auch an die entfernten MCS verteilt werden können. Hierdurch kann die Leitungsauslastung reduziert werden, da keine duplizierten Daten mehr über eine Leitung transportiert werden müssen. Der primäre MCS behandelt und verwaltet die entfernten MCS identisch mit den lokalen Teilnehmer und die entfernten MCS bauen eine Verbindung zum primären MCS genauso auf wie lokale Teilnehmer.

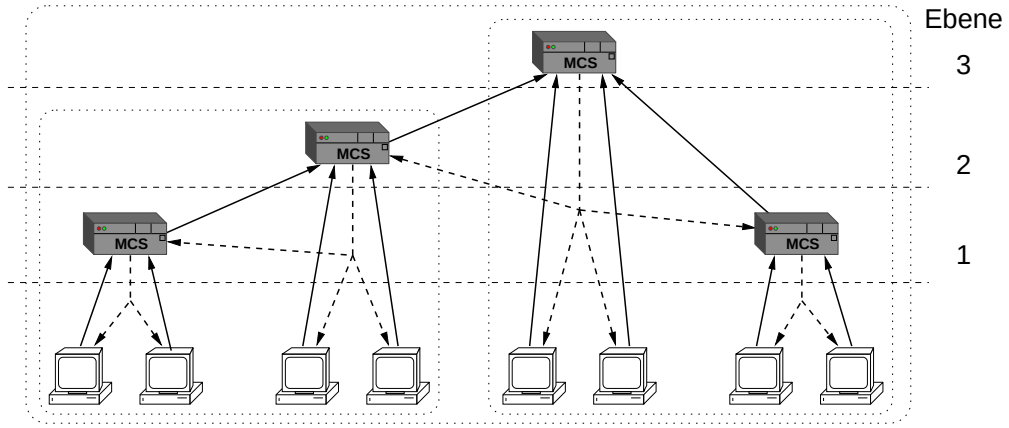


Abbildung 6.9.: Schema bei mehreren Hierarchieebenen.

Bei mehr als zwei Hierarchieebenen ergibt sich eine etwas andere Ordnung, als in Abbildung 6.8(b) dargestellt. Ein primärer MCS für die nächst höhere Ebene wird aus den primären MCS der unteren Ebene ausgewählt. Abbildung 6.9 zeigt das für drei Hierarchieebenen. Die primären MCS übernehmen Aufgaben auf mehreren Hierarchieebenen. Dem primären MCS auf Ebene 3 ist ein lokaler MCS auf Ebene 1 zugeordnet und ein MCS auf Ebene 2. Dem MCS auf Ebene 2 ist noch einen weiterer MCS auf Ebene 1 zugeordnet, er ist also primärer MCS für Ebene 2, hat aber einen weiteren primären MCS auf einer höheren Ebene.

Verwaltung und Datentransfer

Für das Verständnis des hier vorgestellten Ansatzes zur Gruppenkommunikation über ATM-Netzen ist eine strikte Trennung zwischen Verwaltung und Datentransfer entscheidend. Bei beiden wird eine Baumstruktur angewendet, die aber jeweils anderen Zwecken dient und andere Daten transportiert. Um die Unterschiede zwischen den Kommunikationsstrukturen der Verwaltung und beim Datentransfer hervorzuheben, werden in einem Beispiel beide Strukturen gegenübergestellt.

Das Beispiel in Abbildung 6.10 besteht aus vier Teilen. Das zugrunde liegende ATM-Netz ist in Abbildung 6.10(a) dargestellt. Das Netz besteht aus lokalen Teilnetzen, die

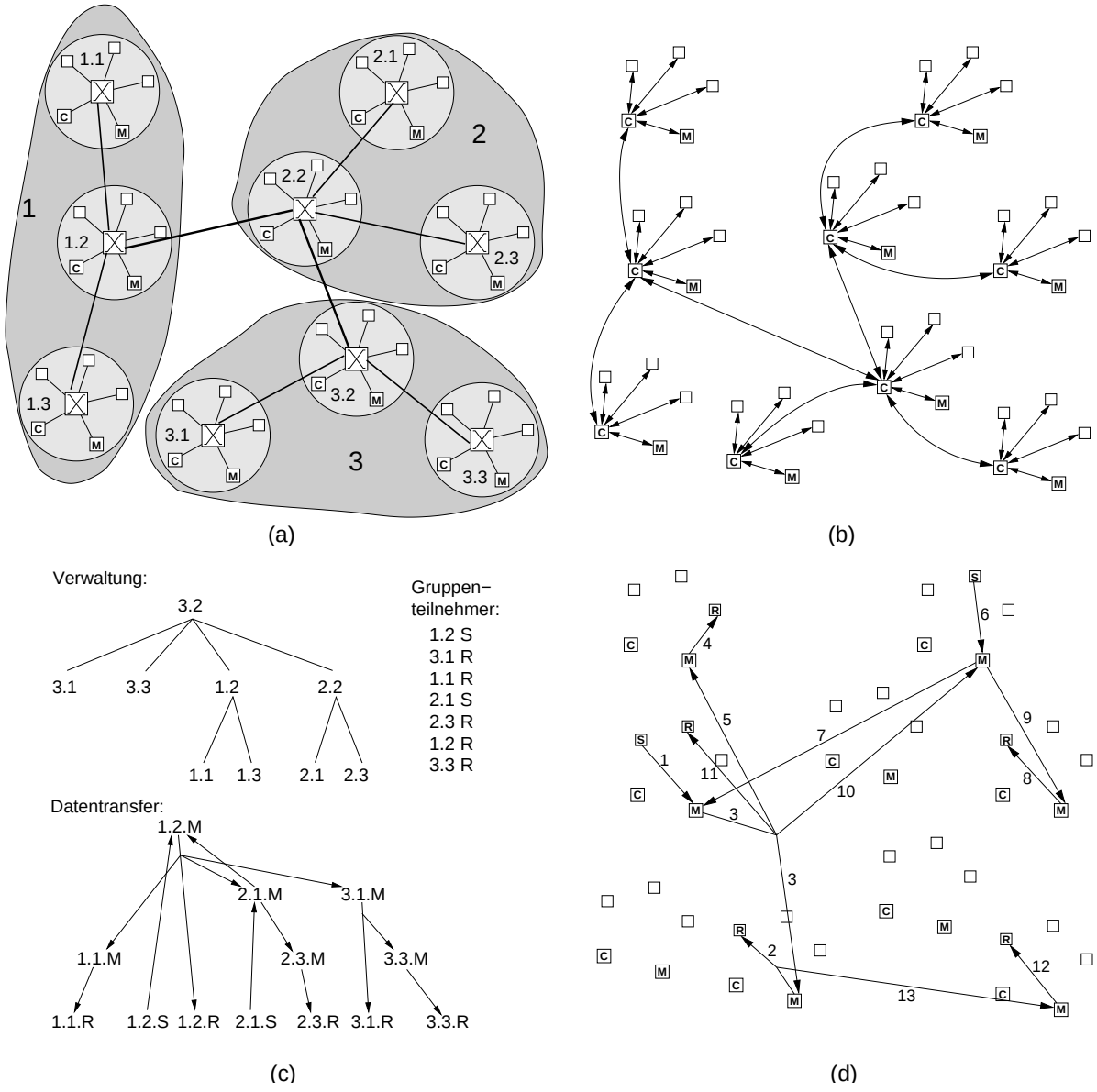


Abbildung 6.10.: Beispiel für die Kommunikationsstrukturen in Verwaltung und Datentransfer.

eine sternförmige Struktur haben und vereinfacht immer aus drei Endsystemen, einem Controller (C) und einem MCS (M) bestehen. Die Teilnetze sind in zwei Stufen organisiert und untereinander verbunden. Diese Strukturierung ist typisch für das PNNI Routing von ATM. Zusätzlich sind die lokalen Teilnetze in der Abbildung mit zweistelligen Adressen versehen. Diese Adressen orientieren sich dabei an die Aufteilung der Teilnetze.

Die nächste Abbildung 6.10(b) stellt die Gruppenverwaltungsstruktur mit den zugehörigen Signalisierungsverbindungen dar. Die ATM-Leitungen, die die Komponenten verbinden, und die Einfärbung der PNNI-Hierarchie sind aus Gründen der Übersichtlichkeit nicht mehr eingezeichnet. In jedem lokalen Teilnetz sind alle Endsysteme und der MCS mit dem Controller verbunden. Die Verbindungen zwischen den Controllern stellen die Hierarchie der Verwaltung dar. Der Controller in Teilnetz 3.2 ist dabei der Wurzelknoten für die Verwaltung.

Die Verwaltungshierarchie ist noch einmal in Abbildung 6.10(c) oben dargestellt. Wie gut zu erkennen ist, ist der Verwaltungsbaum nicht vollkommen ausgeglichen, was aber für die eigentliche Funktionalität keine notwendige Bedingung darstellt. Neben dem Verwaltungsbaum steht eine Liste von Teilnehmeranmeldungen für eine Gruppe. Bei den Teilnehmeranmeldungen ist in erster Linie die Reihenfolge von Bedeutung, da hierdurch indirekt die primären MCS in der Gruppe festgelegt werden (siehe Unterkapitel 6.2.2). Daneben ist natürlich von Bedeutung, ob ein Teilnehmer als Sender (S) oder als Empfänger (R) der Gruppe beitritt. Die durch die Gruppenbeitritte entstehende hierarchische Struktur für den Datentransfer ist in 6.10(c) unten dargestellt. Die Darstellung ist optimiert in bezug auf eine übersichtliche Anordnung des entstandenen Baums für den Datentransfer.

Wie der Baum sich an der Netzwerktopologie orientiert, zeigt Abbildung 6.10(d). Alle Teilnehmer in der Gruppe haben eine Verbindung zu oder von den MCS. Die MCS sind wiederum über die primären MCS der jeweils höheren Ebene untereinander verbunden. An jeder ATM-Verbindung steht eine Nummer, die die Reihenfolge angibt, in der die Verbindungen etabliert worden sind. An der Darstellung kann die Berücksichtigung der räumlichen Organisation beim Aufbau des Baums zum Datentransfer ebenfalls gut erkannt werden.

Paketreflexionen

Bei dem hierarchischen MCS-Schema ist das Problem der Paketreflexionen (siehe auch Unterkapitel 2.4, ab Seite 23) zwischen den Hierarchieebenen ebenfalls ein zu beachtendes Problem. Hierbei ist vor allem wichtig, dass nur ein primärer MCS auf der gleichen ATM-Verbindung sowohl die untergeordneten MCS als auch die lokalen Teilnehmer mit Daten versorgen kann. Eine Nichtbeachtung dieser Regel zeigt Abbildung 6.11(a). Die Kleinbuchstaben in der Abbildung stellen die Pakete des Senders mit dem entsprechenden Großbuchstaben dar. Der untere MCS M1 sendet die Daten auf einer Punkt-zu-Mehrpunkt-Verbindung sowohl an das lokale Endsystem als auch an den primären MCS. Zu den empfangenen Daten zählen aber ebenso die Datenpakete vom primären MCS, wodurch ein Zyklus entsteht und immer mehr reflektierte Datenpakete im Netzwerk ent-

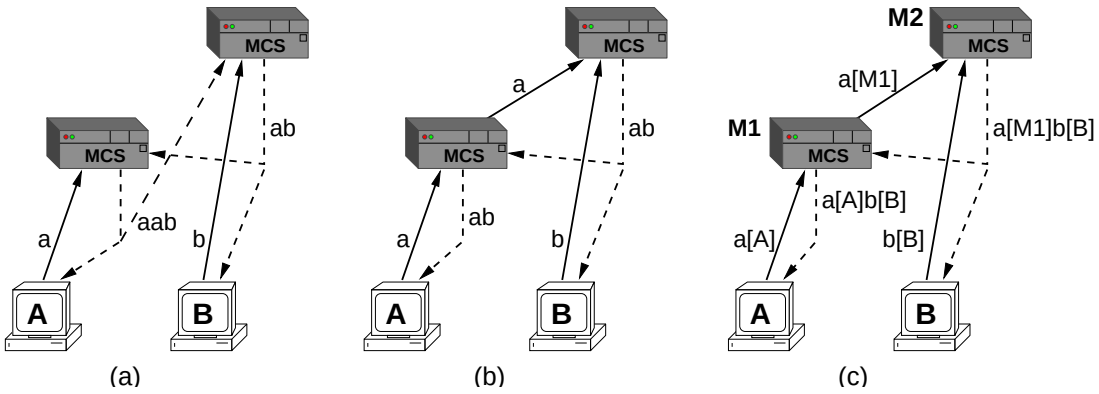


Abbildung 6.11.: Paketreflexionen: (a) Zyklischer Datenverkehr bei Nichtbeachtung der Hierarchieebene, (b) erwünschte Paketverteilung auf den ATM-Verbindungen, (c) Label zur Erkennung von Paketreflexionen.

stehen. Dieser Zyklus kann durch eine Trennung von unteren und höheren Teilnehmern vermieden werden, wie in Abbildung 6.11(b) dargestellt.

Der untere MCS sendet die Datenpakete seiner lokalen Teilnehmer auf einer separaten Verbindung zum primären MCS. Hierbei sind aber weiterhin zwei Faktoren zu beachten:

1. Ein MCS muss lokale Datenpakete erkennen können und nur diese an den höheren MCS weiterleiten.

2. Ein MCS muss bei allen vom höheren MCS empfangenen Datenpaketen, die reflektierten Datenpakete erkennen und verwerfen.

Der erste Punkt, das Erkennen lokaler Datenpakete, kann gelöst werden, indem die Absenderadresse eines Datenpaketes in der MCS-Datenstruktur verifiziert wird. Ist der Absender des Datenpaketes im MCS eingetragen, so ist es ein lokaler Teilnehmer. Für den zweiten Faktor wird das Label-Feld im Datenpaketkopf (Tabelle 4.6, Seite 61) eingesetzt. Das Label wird immer dann mit der eigenen Adresse gefüllt, wenn das Datenpaket zu einem höheren System geschickt wird. Das bedeutet, jedes Endsystem setzt das Label mit der eigenen Adresse und im MCS wird das Label gesetzt, wenn das Datenpaket zum primären MCS gesendet wird. Abbildung 6.11(c) zeigt diese Vorgehensweise. Die Buchstaben in eckigen Klammern stellen die Absenderadresse im Labelfeld dar. Endsystem A setzt seine Adresse als Label im Datenpaket, MCS M1 schickt das Paket zum primären MCS M2 weiter und überschreibt dabei das Labelfeld mit der eigenen Adresse. Der primäre MCS M2 lässt das Labelfeld hingegen unverändert, da die Daten zu lokalen Teilnehmern bzw. zum unteren MCS gesendet werden. Jetzt erkennen MCS M1 und Endsystem B ihre eigenen Datenpakete durch einen Vergleich des Labelfeldes mit der eigenen Adresse und können diese Datenpakete verwerfen. MCS M1 versendet also nur noch das ‚lokale' Paket a[A] für evtl. weitere angeschlossene Endsysteme. A erkennt sein eigenes Paket und kann es verwerfen.

6.2.2. Signalisierung

Für die Etablierung der ATM-Verbindungen nach dem im vorherigen Unterkapitel gezeigten Schema ist eine Signalisierungsunterstützung notwendig. Hierzu wird ein verteilter Ansatz eingesetzt, der auf dem in Unterkapitel 6.1 vorgestellten Konzept zur Gruppenverwaltung aufbaut und dieses erweitert.

Das Signalisierungsprotokoll bildet die Basis für die Etablierung von ATM-Verbindungen für die Gruppenkommunikation. Die Signalisierung wird immer durch eine An- oder Abmeldung von einem Gruppenteilnehmer (Sender oder Empfänger) initiiert. Das Signalisierungsprotokoll ist dabei in zwei Teile aufgeteilt: Signalisierung zwischen Endsystem, MCS und Controller und Signalisierung zwischen Controllern. Der erste Teil, die Signalisierung zwischen Endsystem, Controller und MCS ist identisch mit der Signalisierung aus Unterkapitel 6.1.2. Hier werden nur Nachrichten zur An- und Abmeldung von Teilnehmern ausgetauscht. Der zweite Teil, die Signalisierung zwischen den Controllern, behandelt den eigentlichen Kern des Signalisierungsprotokolls, der eine Gruppenkommunikation über lokale Netzwerkgrenzen hinaus ermöglicht.

Das im Folgenden beschriebene Signalisierungsprotokoll behandelt die reguläre Signalisierung. Ausfälle oder Störungen werden nicht mit in Betracht gezogen. Alle Nachrichten werden mit dem in Unterkapitel 4.2.2, Seite 61 vorgestellten Nachrichtentransportprotokoll versendet, das einen zuverlässigen Transportdienst anbietet. Zunächst wird beschrieben, nach welchem Kriterium ein primärer MCS ausgewählt wird. Danach folgt eine Erläuterung des Systemverhaltens und über die Eigenschaften des Signalisierungsprotokolls. Am Ende wird anhand eines größeren Beispiels der Ablauf des Signalisierungsprotokolls demonstriert.

Wahl des primären MCS

Bei dem in Unterkapitel 6.2.1 beschriebenen Gruppenkommunikationsschema wird die hierarchische Kommunikation mit dem Einsatz von primären MCS ermöglicht. Es ist hingegen nicht beschrieben worden, aufgrund welcher Entscheidung ein MCS zu einem primären MCS wird. Für diese Entscheidung wird zunächst einmal das einfachste Kriterium angewendet: Der MCS, welcher sich als erster für eine Gruppe anmeldet, wird der primäre MCS für diese Gruppe in der jeweiligen Hierarchieebene. Dieses Kriterium ist nicht in jedem Fall optimal und kann zu einer ungünstigen Verbindungswahl führen. Das ist möglich, da der MCS ohne Berücksichtigung der Netzwerktopologie gewählt wird. Ein Mechanismus, der eine bessere Wahlmöglichkeit eines MCS erreicht, wird im nächsten Unterkapitel 6.3 beschrieben. Für das eigentliche Signalisierungsprotokoll ist die Wahl des primären MCS aber nicht von Bedeutung.

Systemverhalten

Zur Charakterisierung des Signalisierungsprotokolls werden hier zuerst einmal die wichtigsten Eigenschaften und Merkmale zusammengefasst erläutert:

- Es wird strikt zwischen beteiligten Sendern und Empfängern unterschieden. Ein

Endsystem, das sowohl Sender als auch Empfänger in einer Gruppe ist, wird wie zwei Endsysteme betrachtet.

- Jede Gruppe wird separat behandelt, es gibt keinerlei Interaktionen zwischen unterschiedlichen Gruppen. Das vereinfacht das Protokoll, denn das Problem der Signalisierung muss nur einmal für eine Gruppe gelöst werden.

- Die Vergabe von Gruppenadressen und die Bekanntgabe von Gruppenadressen bei den Empfängern wird nicht behandelt. Es wird davon ausgegangen, dass ein Announcement-Protokoll wie z. B. SAP [59] bei den Teilnehmern eingesetzt wird, das für die Bekanntgabe von Gruppenadressen zuständig ist.

- Die Sender/Empfänger-Information wird genutzt, um Datenpakete von Sendern in Richtung der Empfänger zu leiten. Ein Sender kennt nur beteiligte Empfänger und keine weiteren Sender, ein Empfänger hat hingegen keine Kenntnis über seine Sender. Hierbei können nicht nur Endsysteme sondern auch MCS die Sender oder Empfänger sein.

- Primäre MCS repräsentieren stellvertretend die Empfänger von anderen Hierarchieebenen. Sobald ein Sender vorhanden ist, ist ein MCS immer auch Empfänger für die Gruppe, damit er die Daten an die untergeordneten Empfänger weitergeben kann.

- Der MCS, der sich zuerst für eine Gruppe anmeldet, wird der primäre MCS in der unteren und in den höheren Hierarchieebenen. Jeder MCS hat das Bestreben, den Zustand des primären MCS in höheren Hierarchieebenen anzunehmen. Ist in einer höheren Hierarchieebene schon ein primärer MCS vorhanden, so wird der neue MCS bei einer 'Kollision' informiert, dass er nur primärer MCS für seine aktuelle Hierarchieebene ist.

- Hat ein primärer MCS keine lokalen Gruppenteilnehmer mehr, so wäre es wünschenswert, dass er auch seine Funktion als primärer MCS abgibt und keine weiteren Gruppendaten mehr erhält. Dieses Verhalten lässt das beschriebene Signalisierungsprotokolls nicht zu. Eine Erweiterung, die dieses Verhalten in Zusammenhang mit einer besseren Wahl des primären MCS kombiniert, ist in Unterkapitel 6.3 beschreiben.

- Gegenüber der in Unterkapitel 6.1 beschriebenen hierarchischen Gruppenverwaltung werden Nachrichten nicht nur von Controllern in unteren Hierarchieebenen zu Controllern in höheren Hierarchieebenen gesendet. Für die Etablierung von ATM-Verbindungen zum Datentransfer findet ein Nachrichtenaustausch in beiden Richtungen im Verwaltungsbaum statt. Nachrichten von unteren Hierarchieebenen werden benutzt, um den Informationsstand in den höheren Hierarchieebenen zu aktualisieren. Nachrichten, die von höheren Hierarchieebenen nach unten gesendet werden, lösen hingegen konkrete Aktionen (Auf- und Abbau von ATM-Verbindungen) bei den beteiligten Komponenten aus.

- Die Baumstruktur der Gruppenverwaltung ist nicht zu verwechseln mit der Baumstruktur für den Datentransfer innerhalb einer Gruppe. Der Verwaltungsbaum wird durch eine manuelle (externe) Konfiguration festgelegt und ist im Prinzip eine statische Struktur. Der Baum für den Datentransfer in einer Gruppe entsteht dagegen durch die Anmeldereihenfolge der Endsysteme. Die MCS, bei denen sich die ersten Teilnehmer für eine Gruppe anmelden, werden die primären MCS auf den einzelnen Hierarchieebenen.

- Die Baumhöhe der Gruppenverwaltung limitiert auch die maximale Baumhöhe des Datentransferbaums. Die Anzahl der Zwischensysteme kann nie größer als die Baumhöhe sein, da jeder MCS immer genau einem Controller zugeordnet ist. Der Datentransferbaum kann aber auch eine wesentlich geringere Baumhöhe haben, das ist abhängig von der Gruppengröße.

- Nur die Detailschicht im Controller ist für den konkreten Auf- und Abbau von ATM-Verbindungen verantwortlich. Bei An- oder Abmeldung eines Teilnehmers signalisiert die Detailschicht dem MCS die notwendigen Änderungen, woraufhin der MCS die entsprechenden ATM-Verbindungen auf- oder abbaut.

- Die Aggregationsschicht im Controller dient hauptsächlich der Kommunikation mit der nächst höheren Hierarchieebene. Werden der Aggregationsschicht entfernte Gruppenteilnehmer von einem höheren Controller mitgeteilt, so leitet die Aggregationsschicht die Nachrichten dieser Teilnehmer an die Detailschicht weiter.

- Die Verteilung der Datenpakete innerhalb einer Gruppe hat Ähnlichkeiten mit Core Based Trees, wobei der höchste primäre MCS den Core bildet und die Gruppen separate Spannbäume benutzen.

- Es gibt immer nur einen Baum pro Gruppe. Das kann zu Verkehrskonzentrationen und Engpässen führen, da alle Sender einer Gruppe sich dieselben MCS für die Datenweiterleitung teilen. In Unterkapitel 6.3 wird eine Lösung vorgestellt, die das Verwenden mehrerer Bäume pro Gruppe ermöglicht.

Das Signalisierungsprotokoll wird zuerst an einem Beispiel demonstriert. Hierbei werden die grundlegenden Mechanismen und der Nachrichtenaustausch beschrieben, die ein hierarchisches Gruppenkommunikationsschema aufbauen.

Beispiel:

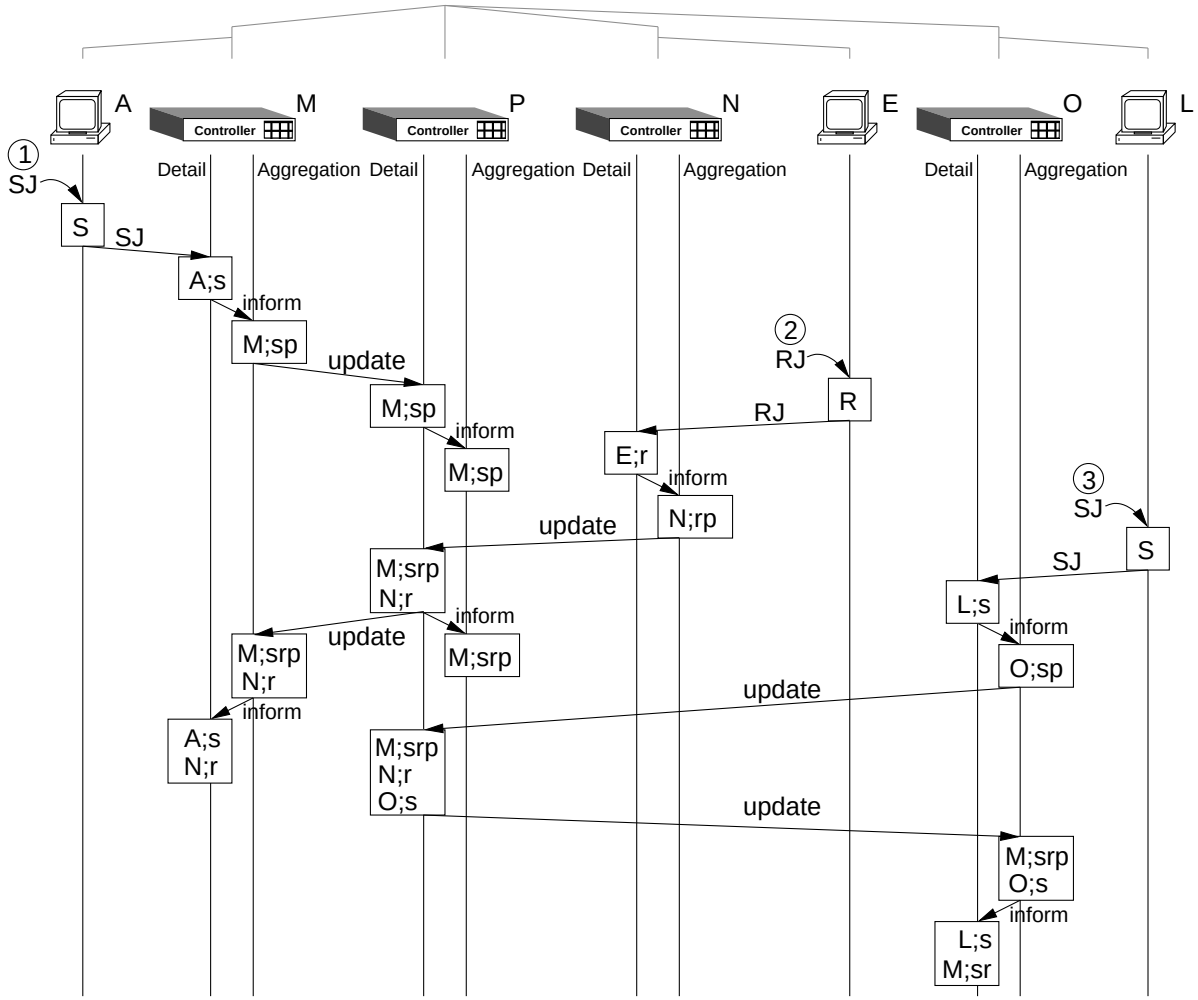
Das Beispiel orientiert sich an dem Szenario in Abbildung 6.1 auf Seite 87. Es gibt für jedes lokale ATM-Netz einen Controller und einen MCS, der dem Controller zugeordnet ist. Zur Vereinfachung der Darstellung und Erhöhung der Übersichtlichkeit werden MCS und Controller zusammengefasst und im Weiteren nur noch als ein Controller dargestellt. Wenn ein Controller spezifiziert wird, so ist damit auch immer der zugehörige MCS bezeichnet. Im Folgenden wird daher auch der Begriff primärer Controller verwendet. Eine weitere Vereinfachung in diesem Beispiel ist die Beschränkung auf eine

Gruppe. Alle Aktionen und Nachrichten beziehen sich nur auf eine Gruppe. Daher werden die Gruppenadressen in diesem Beispiel nicht dargestellt, obwohl jede Nachricht und jeder Zustand gruppenspezifisch ist. Das Signalisierungsprotokoll behandelt jede Gruppe separat.

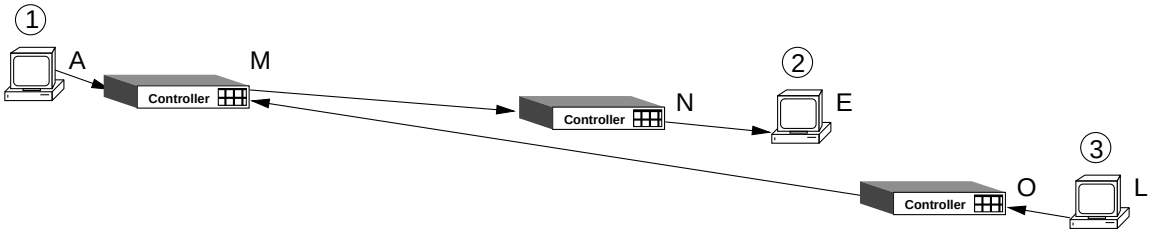
Abbildung 6.12 zeigt das Beispiel und die Signalisierung zwischen den Controllern. Ganz oben in Abbildung 6.12 ist die hierarchische Gliederung der Controller angedeutet, wie sie bei der Gruppenverwaltung etabliert worden ist. Der untere Teil zeigt die Reihenfolge, in der die Teilnehmer beitreten, und wie die ATM-Verbindungen zwischen den Komponenten aufgebaut werden. Der mittlere Teil stellt den Ablauf des Signalisierungsprotokolls als Weg-Zeit-Diagramm dar. In diesem Beispiel tritt zuerst Endsystem A als Sender der Gruppe bei. Als nächstes kommt Endsystem E als Empfänger hinzu und am Ende tritt Endsystem L als weiterer Sender der Gruppe bei. Diese Reihenfolge bewirkt, dass der Controller M in allen Hierarchieebenen primärer Controller für die Gruppe wird und die Controller N und O lokale primäre Controller werden. Im Weg-Zeit-Diagramm ist bei den Controllern für die Detail- und die Aggregationsschicht jeweils eine separate Zeitlinie vorhanden. Damit wird ergänzend der interne Nachrichtenaustausch im Controller verdeutlicht.

Der Beitritt von Endsystem A bewirkt, dass eine Senderbeitrittsnachricht (Sender-Join=SJ) an den Controller M gesendet wird. Das Endsystem A ist der erste Gruppenteilnehmer im Controller, woraufhin eine neue Verwaltungsstruktur für die Gruppe im Controller angelegt wird. Das Endsystem A wird in der Detailschicht für die Gruppe als Sender eingetragen (A;s)und die Aggregationsschicht wird über die neue Gruppe und den Sendestatus informiert. Da die Gruppe in der Aggregationsschicht neu ist, erklärt sich der Controller M zum primären Controller und Sender (M;sp = **s**ender und **p**rimary) für die Gruppe. Dieser Status wird an den nächst höheren Controller P weitergeleitet. Der Controller P trägt den Controller M als primären Controller für die Gruppe in die Detail- und Aggregationsschicht ein. Es werden keine weiteren Aktionen ausgeführt, da keine anderen Gruppenmitglieder vorhanden sind und P der höchste Controller ist.

Das nächste Ereignis ist der Beitritt von Endsystem E als Empfänger für die Gruppe. Das Vorgehen ist analog zu Endsystem A, eine Empfängerbeitrittsnachricht (Receiver-Join=RJ) wird an den Controller N gesendet. Das Endsystem wird in die Detailschicht aufgenommen und in der Aggregationsschicht erklärt sich der Controller N zum primären Controller und Empfänger für die Gruppe (N;rp = **r**eceiver und **p**rimary). Dieser Zustand wird an den Controller P gesendet, wo aber bereits der Controller M als primärer Controller für diese Gruppe registriert ist. Dem Controller N wird der Status als primärer Controller aberkannt und der Controller M wird als Empfänger markiert, da er die Gruppe nach außen hin repräsentiert und die Eigenschaften aller Gruppenteilnehmer zusammenfasst. Diese Änderung wird jetzt an den Controller M gesendet. Der Controller M fügt N als Empfänger in der Aggregationsschicht hinzu und informiert die Detailschicht über den neuen Empfänger N. Der Controller N wird in der Detailschicht wie ein lokales System behandelt und es wird eine ATM-Datenverbindung zu Controller N aufgebaut. Der Controller N wird nicht benachrichtigt, da er als Empfänger nur ein 'passives' System ist, das Datenpakete entgegennimmt. Solange er lokal keine Sender hat, ist es nicht wichtig, dass er die Adresse des zuständigen primären Controllers kennt.



(a) Signalisierung



(b) Datenfluss

Abbildung 6.12.: Signalisierung (a) und Datenfluss (b) zwischen Controllern.

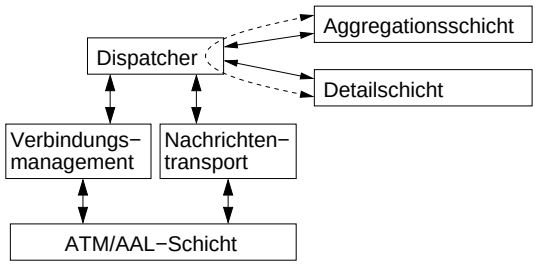


Abbildung 6.13.: Aufbau des Controllers.

Das letzte Ereignis ist der Beitritt von Endsystem L als weiterer Sender in der Gruppe. Im Controller O werden die gleichen Aktionen ausgeführt wie zu Beginn bei Controller M. Im Controller P wird der Controller O als Sender hinzugefügt. Bei den anderen registrierten Controllern M und N ändert sich dadurch nichts. Im Unterschied zum vorhergehenden Empfängerbeitritt muss dem Controller O der korrekte primäre Controller für die Gruppe mitgeteilt werden, damit die Datenpakete an alle Empfänger verteilt werden können. Controller O erhält die Adresse des primären Controllers M und trägt diesen bei sich in der Aggregations- und Detailschicht ein. In der Detailschicht wird der primäre Controller wie ein lokaler Empfänger behandelt und es wird eine ATM-Datenverbindung zum Controller M aufgebaut.

6.2.3. Datenhaltung und Organisation im Controller

Im vorherigen Unterkapitel ist der Ablauf des Signalisierungsprotokolls zwischen den Controllern im Gesamtzusammenhang erläutert worden. Für eine konkrete Realisierung ist es hingegen wichtig, die Funktionalität der beteiligten Komponenten, also der Controller, zu beschreiben. Zur Funktionalität gehören zum einen Aktionen und zum anderen Daten, auf denen die Aktionen ausgeführt werden. Der Controller arbeitet ereignisorientiert, d. h. es werden im Allgemeinen nur Aktionen ausgeführt, wenn Nachrichtenpakete beim Controller eintreffen. Zusätzlich existieren noch zeitbasierte Aktionen, z. B. die periodische Aktualisierung von Zustandsinformationen, die aber für die eigentliche Nachrichtenverarbeitung nur von geringer Bedeutung sind.

Im Unterkapitel 6.1 auf Abbildung 6.6 ist schon das Zwei-Schichten-Modell des Controllers beschrieben worden. Dort findet die Kommunikation zwischen den Schichten allerdings nur in einer Richtung statt, von der Detailschicht zur Aggregationsschicht. Für die Etablierung des Datentransfers innerhalb einer Gruppe ist es erforderlich, dieses Modell zu erweitern. Abbildung 6.13 zeigt den Aufbau des Controllers. Zwischen der Aggregations- und der Detailschicht ist ein Verteiler (Dispatcher), der eine Reihe von Aufgaben erfüllt:

- Aufbau und Annahme von ATM-Verbindungen zu anderen Controllern, Endsystemen und MCS. Im Dispatcher werden auch die Verbindungen zu anderen Controllern mit dem in Unterkapitel 6.1.1 beschriebenen Algorithmus etabliert.
- Verteilen von ankommenden Nachrichten an die Aggregations- oder Detailschicht.

Anhand der ATM-Verbindung, auf der die Nachricht angekommen ist, wird entschieden, welche Schicht die Nachricht erhält. Alle Nachrichten von Endsystemen, MCS und untergeordneten Controllern gehen an die Detailschicht. Diese Nachrichten kommen alle auf ATM-Verbindungen an, die vom Controller angenommen, aber nicht selbst aufgebaut worden sind. Nachrichten von höheren Controllern gehen an die Aggregationsschicht. Diese Nachrichten kommen immer auf einer ATM-Verbindung an, die der Controller selbst initiiert hat.

- Für die Kommunikation zwischen den Schichten existiert eine Pseudo-Verbindung (gestrichelte Linie zwischen den Schichten in Abbildung 6.13). Der Austausch von Informationen zwischen den Schichten geschieht mittels derselben Nachrichten, wie sie auch für die Kommunikation mit anderen Controllern eingesetzt werden. Durch das Verpacken und Entpacken von Informationen und die zusätzliche Einbeziehung des Dispatchers erfordert dieses Konzept einen zusätzlichen Verarbeitungsaufwand im Controller. Dafür hat diese Vorgehensweise den Vorteil einer kompletten Trennung der beiden Schichten, wodurch die Entwicklung der einzelnen Schichten vereinfacht wird.

Die beiden Schichten im Controller können jetzt auf die wesentlichen Aufgaben beschränkt werden, da alle Teile, die in die Netzwerkkommunikation involviert sind, aus den Schichten ausgelagert werden können. Hinzu kommt, dass beide Schichten viele Gemeinsamkeiten bzgl. des internen Ablaufs und der verwendeten Datenstrukturen haben. Daher wird zunächst ein allgemeines Modell einer Schicht im Controller beschreiben, woraufhin dann die Unterschiede zwischen der Detail- und der Aggregationsschicht erläutert werden.

Allgemeines Modell einer Controller-Schicht

Wie schon erwähnt, arbeitet der Controller ereignisgesteuert und reagiert auf ankommende Nachrichten. Das gleiche Verhalten liegt auch den beiden Schichten im Controller zugrunde. Jede Schicht hat ein mehrstufiges Vorgehensmodell, das in Abbildung 6.14 dargestellt ist.

Trifft eine Nachricht ein, so wird deren Inhalt extrahiert und in den Datenbestand mit aufgenommen. Das kann bedeuten, dass Daten anhand der Nachricht aktualisiert oder dass neue Daten hinzugefügt werden. Im nächsten Schritt wird der Datenbestand (Abbildung 6.15) geprüft und aufgrund der festgestellten Veränderungen werden nach Bedarf neue Nachrichten generiert. Diese generierten Nachrichten werden an den Dispatcher übergeben, der sie dann an andere Controller, MCS oder die jeweils andere Controller-Schicht weitergibt. Im letzten Schritt wird der Datenbestand aufgeräumt, was im Grunde nur darin besteht, nicht mehr benötigte Einträge aus dem Datenbestand zu löschen. Das Vorgehen beim Eintreffen von Nachrichten besteht also immer aus drei Schritten: aktualisieren, auswerten und aufräumen.

Der Datenbestand ist der wesentliche Teil einer Controller-Schicht. Daher ist es wichtig zu wissen, wie die Daten im Controller organisiert sind. Wie schon im Unterkapitel 6.1 beschrieben, werden die einzelnen Gruppen separat voneinander gehandhabt. Das

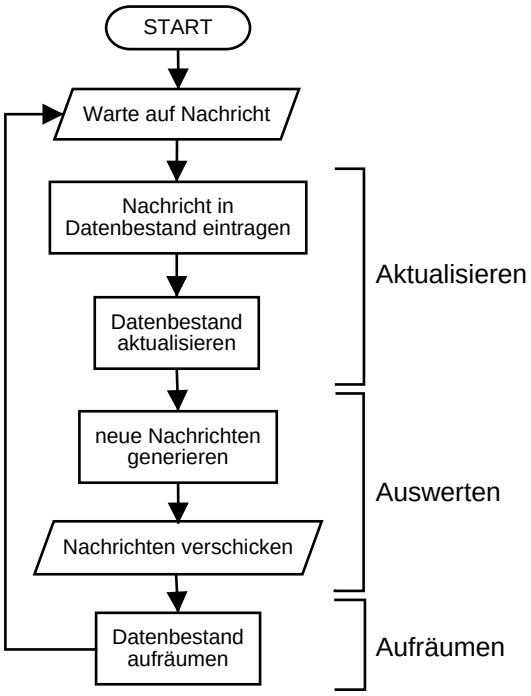


Abbildung 6.14.: Aufbau einer Schicht im Controller.

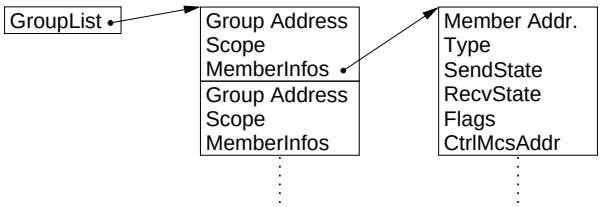


Abbildung 6.15.: Aufbau des Datenbestands in einer Controller-Schicht.

erlaubt eine erste Strukturierung der Daten. Es existiert eine Liste von Gruppen mit den zugehörigen Gruppenadressen, Reichweiten und weiteren Daten. Jede ankommende Nachricht hat eine Gruppenadresse, mit der sie in der Schicht der entsprechenden Gruppe zugeordnet und dort weiterverarbeitet werden kann. Jede Gruppe besteht wiederum aus einer Menge von Teilnehmern, die allerdings verschiedene Eigenschaften und Attribute aufweisen können. Den Aufbau des Datenbestands gibt Abbildung 6.15 wieder. Die Liste der Gruppen ist sortiert und kann anhand der Gruppenadresse durchsucht werden. Mit dem gefundenen Gruppeneintrag können dann weitere Informationen über die Teilnehmer der Gruppe abgefragt werden.

Zu jedem Teilnehmer einer Gruppe wird dessen Adresse gespeichert, der Typ eines Teilnehmers ist entweder ENDSYSTEM oder MCS. Jeder Teilnehmer kann sowohl Sender als auch Empfänger in der Gruppe sein, worüber die Variablen SendState und RecvState Auskunft geben. Eine wichtige Unterscheidung ist noch, in welchem Zustand der Teilnehmer sich dabei gerade befindet: Inaktiv (Idle), Beitreten (Add), Aktiv (Active)

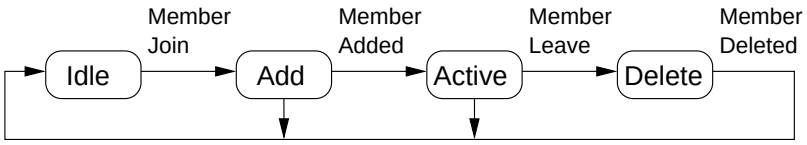


Abbildung 6.16.: Die Zustände eines Teilnehmers und die möglichen Zustandsübergänge.

oder verlassen (Delete). Der Wechsel zwischen diesen Zuständen ist vorgegeben und in Abbildung 6.16 dargestellt. Als Flag ist im Moment nur die Information PRIMARY vorgesehen, die einen MCS als primären MCS kennzeichnet. Das Feld CtrlMcsAddr ist für entfernte MCS notwendig und nur beim Typ MCS gesetzt. Das Feld enthält die Adresse des zuständigen Controllers, über den der entfernte MCS erreicht werden kann. Enthält das Feld eine ungültige Adresse, kennzeichnet dies einen lokalen MCS.

Mit dieser Datenstruktur können alle relevanten Daten gespeichert werden. Die darauf stattfindenden Aktionen gliedern sich, wie oben beschrieben, in drei Schritte: aktualisieren, auswerten und aufräumen. Der erste und letze Schritt sind sehr einfach:

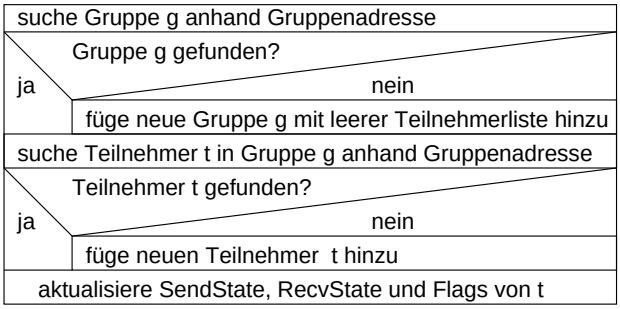
Aktualisieren: Hier wird zuerst der Teilnehmer in der Liste anhand seiner Adresse gesucht. Ist kein Teilnehmer vorhanden, so wird ein neuer Eintrag in die Teilnehmerliste hinzugefügt. Bei dem gefundenen Teilnehmereintrag werden anschließend die Typ- und Zustandsfelder anhand des Nachrichteninhalts (Abbildung 6.17(a)) gesetzt.

Aufräumen: Prüft jeden Teilnehmer in allen Gruppen, ob dieser gelöscht werden kann und löscht gegebenfalls diesen Teilnehmer. Hat eine Gruppe keine weiteren Teilnehmer, wird auch die Gruppe gelöscht (Abbildung 6.17(b)). Die Zustände der Teilnehmer werden überprüft und evtl. neu gesetzt. So werden z. B. Teilnehmer aus dem Zustand Beitreten (Add) in den Zustand Aktiv (Active) gesetzt.

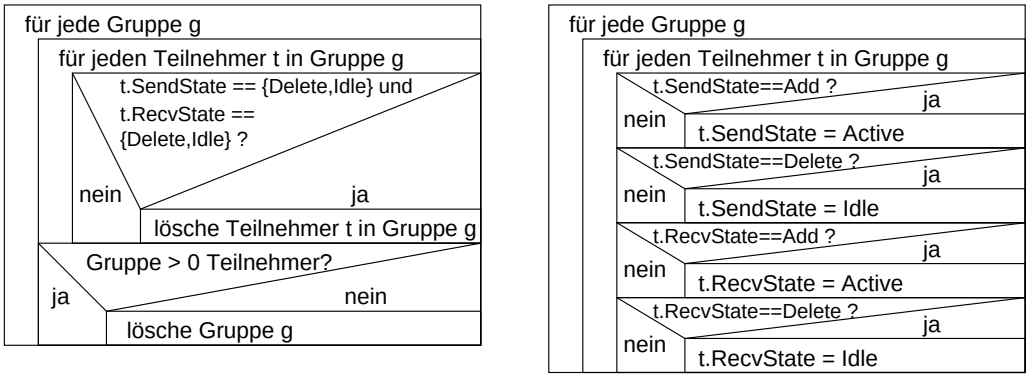
Das Auswerten der Daten ist in der Aggregations- und Detailschicht unterschiedlich implementiert und wird daher im Folgenden getrennt beschreiben. Es werden die wesentlichen Schritte der Auswertung der Datenbestände in den beiden Schichten beschrieben. Viele Details, die die Fehlerbehandlung und Konsistenzerhaltung des Datenbestandes betreffen, sind bei der Beschreibung weggelassen worden.

Detailschicht

In der Detailschicht werden alle lokalen Teilnehmer und untergeordneten MCS behandelt. Bei den MCS ist dabei noch zwischen einem lokalen und einem entfernten MCS zu unterscheiden. In der Detailschicht können die folgenden Nachrichtentypen eintreffen (die Formate sind in Anhang C ab Seite 177 beschrieben): SenderJoin, SenderLeave, ReceiverJoin, ReceiverLeave, Update, Change. Die ersten vier Nachrichten kommen von Endsystemen, die Update-Nachricht von einem unteren Controller und die Change-Nachricht hat die Aggregationsschicht als Ursprung. Es ist aber für die weiteren Verarbeitungsschritte nicht mehr relevant, welche Nachricht eingetroffen ist. Im ersten Schritt bei



(a) Aktualisieren.



(b) Aufräumen.

Abbildung 6.17.: Struktogramme zum Aktualisieren und Aufräumen der Datenbestände in einer Controller-Schicht.

der Nachrichtenverarbeitung, dem Aktualisieren (siehe Abbildung 6.14), werden die Informationen der Nachricht in den Datenbestand eingetragen. Alle Informationen können dann anschließend aus dem Datenbestand, und zwar aus den Zuständen der Teilnehmer (SendState und RecvState) ermittelt werden.

Der erste Schritt ist immer, die Teilnehmer der betreffenden Gruppe zu durchsuchen, und zwar nach einem Teilnehmer, der neu hinzugekommen ist oder gelöscht werden soll (Abbildung 6.18). Je nachdem, ob der Teilnehmer ein Endsystem oder MCS ist, werden unterschiedliche Funktionen ausgeführt.

Ist ein Endsystem der Gruppe neu beigetreten oder wird es die Gruppe verlassen, so wird der lokale MCS aktualisiert. Hat sich am Zustand des lokalen MCS durch den Teilnehmer etwas geändert, so wird eine Nachricht generiert, und an den MCS gesendet. Die MCS-Zustandsänderung erfordert eine weitere Behandlung des MCS. Die Abbildung 6.19 zeigt dieses Vorgehen als Struktogramm der Funktion HandleEndsystem(t) für den Fall einer Senderan- oder -abmeldung. Die Behandlung als Empfänger erfolgt analog und ist nicht dargestellt. Zu bemerken ist noch, dass der lokale MCS immer als primärer MCS markiert wird. Dies wird, falls es bereits einen anderen MCS gibt, in der Funktion

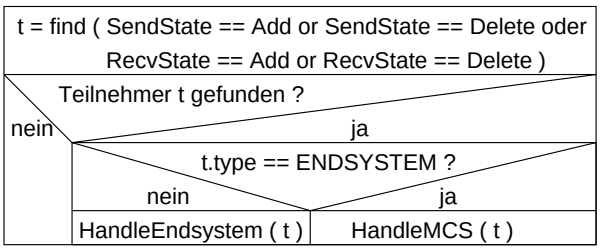


Abbildung 6.18.: Struktogramm zum Finden des geänderten Teilnehmers in der Detailschicht.

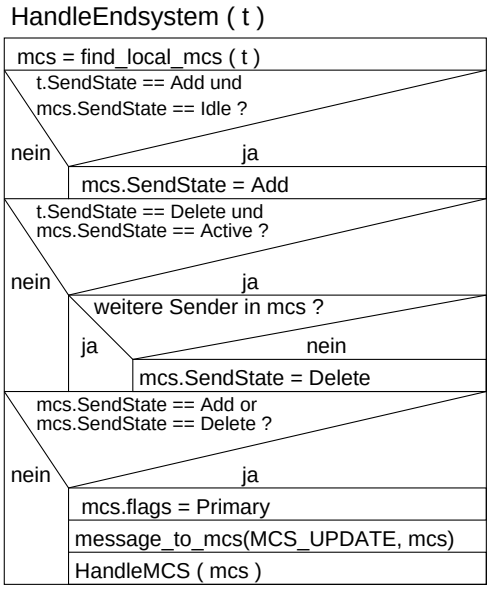


Abbildung 6.19.: Struktogramm zum Anmelden eines Endsystems in der Detailschicht.

HandleMCS wieder zurückgenommen.

Bei einer Änderung eines in der Detailschicht angemeldeten MCS oder beim Eintreffen einer Update-Nachricht von einem unteren Controller wird immer die Funktion HandleMCS aufgerufen (Abbildung 6.20). Hier wird zuerst geprüft, ob der aktuelle MCS primärer MCS ist oder ob es bereits einen anderen primären MCS gibt. Darauf teilt sich die Behandlung des aktuellen MCS auf, je nachdem, ob er primärer MCS ist (HandleMCSPrimary) oder nicht (HandleMCSNormal). Ist der aktuelle MCS der primäre MCS der Detailschicht, so kann dieser die Gruppe nicht verlassen, und es wird eine Nachricht an die Aggregationsschicht mit der Änderung in der Gruppe geschickt. Ist der aktuelle MCS kein primärer MCS, so wird in jedem Fall die Aggregationsschicht informiert und ist der aktuelle MCS als Sender neu hinzugekommen, so wird eine Nachricht an den MCS gesendet, die einen Verbindungsaufbau zum primären MCS einleitet.

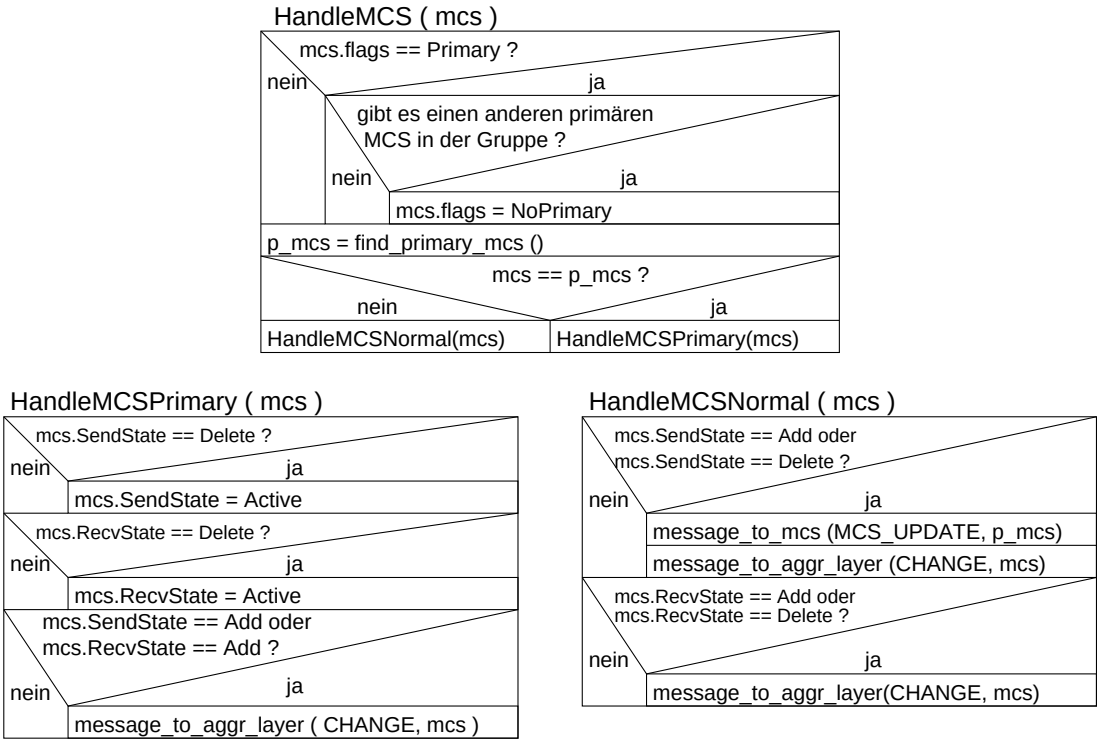


Abbildung 6.20.: Struktogramm zum Anmelden eines MCS in der Detailschicht.

Aggregationsschicht

Die Aggregationsschicht ist für die Kommunikation mit einem Controller aus einer höheren Hierarchieebene zuständig. Existiert kein weiterer Controller in einer höheren Hierarchieebene, so hat die Aggregationsschicht keine Aufgaben zu erfüllen. Die in der Aggregationsschicht gespeicherten Daten stellen gewissermaßen die Sicht des Controllers und dessen untergeordnete Komponenten nach außen dar. Des Weiteren ist in der Aggregationsschicht immer der lokale primäre MCS des untergeordneten Baums enthalten. Dieser lokale MCS steht stellvertretend für alle anderen Systeme im untergeordneten Baum. Zusätzlich sind in der Aggregationsschicht noch entfernte MCS gespeichert, die mit dem lokalen primären MCS kommunizieren.

In der Aggregationsschicht können nur die Nachrichtentypen Update und Change eintreffen (die Formate sind in Anhang C ab Seite 177 beschrieben). Die Update-Nachricht kommt dabei immer vom höheren Controller und die Change-Nachricht von der Detailschicht. In der Aggregationsschicht werden nur MCS gespeichert. Das vereinfacht die Auswertung der Daten, da keine Endsysteme berücksichtigt werden müssen. Die Hauptaufgabe der Aggregationsschicht ist die Bestimmung des primären MCS und anhand dieser Bestimmung die Koordinierung der Nachrichten, was im Folgenden beschrieben wird.

Ist eine Nachricht eingetroffen, wird zuerst der betreffende Teilnehmer im Datenbestand gesucht. Das geschieht genauso, wie in der Detailschicht und ist im Struktogramm

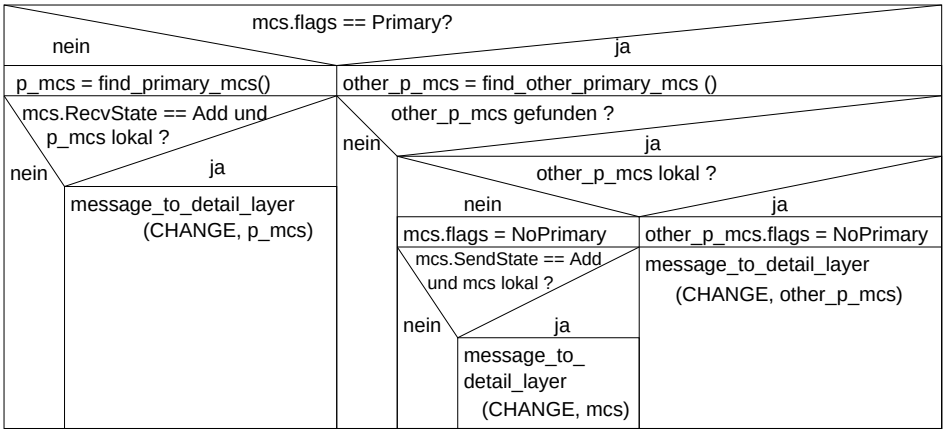


Abbildung 6.21.: Struktogramm der Aggregationsschicht.

in Abbildung 6.18 dargestellt. Daraufhin wird geprüft, ob der aktuelle MCS ein primärer MCS ist und weitere Aktionen eingeleitet (Abbildung 6.21). Ist der aktuelle MCS kein primärer MCS, wird der primäre MCS über den aktuellen MCS informiert. Im anderen Fall muss geprüft werden, ob es nicht bereits einen anderen primären MCS gibt und eine Kollision vorliegt. Gibt es keinen anderen primären MCS, ist alles in Ordnung. Im anderen Fall wird weiter unterschieden, welcher der beiden primären MCS lokal ist. Der lokale MCS bekommt dann den Status als primärer MCS aberkannt und der entfernte MCS bleibt primärer MCS.

Dieses Vorgehen impliziert, dass ein entfernter MCS immer über den höheren Controller erreicht wird (ansonsten wäre er ja ein lokaler MCS). Benennt der höhere Controller einen MCS zum primären MCS, so hat das immer Vorrang vor der eigenen MCS-Auswahl des Controllers, und der gewählte lokale primäre MCS im Controller muss korrigiert werden. Dies beschreibt das Vorgehen im Struktogramm in Abbildung 6.21, in dem die Unterscheidung zwischen lokalen und entfernten MCS zur Bestimmung des primären MCS genutzt wird.

6.2.4. Zusammenfassung

Mit dem hier in Unterkapitel 6.2 vorgestellten Schema zur Gruppenkommunikation über ATM-Weitverkehrsnetze ist es möglich, Gruppenkommunikation über lokale Netze hinaus zu etablieren. Durch die hierarchische Strukturierung können Gruppenteilnehmer zusammengefasst und als ein abstrakter Teilnehmer nach außen hin repräsentiert werden. Das erlaubt die Unterstützung von großen Gruppen, ohne dabei an Restriktionen der ATM-Schicht gebunden zu sein. Die Restriktionen betreffen dabei hauptsächlich die Begrenzung der Anzahl an Endsystemen bei Punkt-zu-Mehrpunkt-Verbindungen und die Beschränkung von ATM-Verbindungen bei ATM-Netzwerkkarten, die bei den bisherigen Konzepten die Skalierbarkeit begrenzt hat (Unterkapitel 3.2.1, ab Seite 32).

Das gewählte hierarchische Konzept ermöglicht darüber hinaus eine Reduzierung bei der Gruppenverwaltung, da hierbei lokale Systeme zusammengefasst und als ein System

nach außen repräsentiert werden. Hierdurch wird der Signalisierungsaufwand und die zu verwaltende Datenmenge in den Zwischensystemen verringert.

Das in diesem Unterkapitel vorgestellte Verfahren löst das Problem der skalierbaren Gruppenkommunikation über ATM-Weitverkehrsnetzen in den wesentlichen Problembereichen. Es sind allerdings noch einige Aufgaben vorhanden, die bisher nur unzureichend gelöst worden sind:

- Das Konzept behandelt hauptsächlich die Anmeldung von Gruppenteilnehmern, die in die Gruppe mit aufgenommen werden müssen. Gruppenteilnehmer können sich zwar auch von der Gruppe abmelden, aber anschließend kann der zugehörige primäre MCS die Gruppe nicht mehr verlassen.

- Die Auswahl des primären MCS wird nur durch die zeitliche Reihenfolge (First Come First Serve) bestimmt. Andere Kriterien, wie z. B. die MCS-Belastung (Unterkapitel 5.3.3, Seite 70) finden keinerlei Berücksichtigung bei der Wahl des primären MCS.

- Es wird nur ein Baum pro Gruppe unterstützt. Das kann zu Verkehrskonzentrationen und daraus resultierenden erhöhten Verzögerungen und Datenverlusten bei den MCS führen.

- Der Einsatz von mehreren lokal vorhandenen MCS ist nicht möglich. In diesem Konzept kann nur ein einziger lokaler MCS für die lokalen Gruppenteilnehmer eingesetzt werden.

Diese Probleme greift das folgende Unterkapitel 6.3 auf, wo zwei Lösungsansätze präsentiert werden, die den vorgestellten Ansatz dahingehend erweitern.

6.3. Erweiterungen für eine verbesserte Lastverteilung

In diesem Unterkapitel werden zwei Ansätze[60] vorgestellt, die bisher noch unzureichend gelöste Probleme bei SkaGAN behandeln. Hierzu zählen:

1. Einmal gewählte primäre MCS können die Gruppe nicht mehr verlassen. Das hat zur Folge, dass MCS weiterhin Daten einer Gruppe weiterleiten müssen, auch wenn sie keine lokalen Teilnehmer mehr besitzen.

2. Ein einmal etablierter Baum ist statisch, d. h. er kann sich nicht an topologische und geografische Änderungen durch Teilnehmerwechsel anpassen. Der Baum kann dadurch degenerieren und die Netzwerkbelastung ansteigen.

3. Alle Sender einer Gruppe teilen sich einen Baum für den Datentransfer. Das kann zu einer ungleichen Belastung des ATM-Netzes führen, da nur wenige Komponenten und Netzwerkverbindungen stark belastet werden und andere Komponenten hingegen nicht verwendet werden.

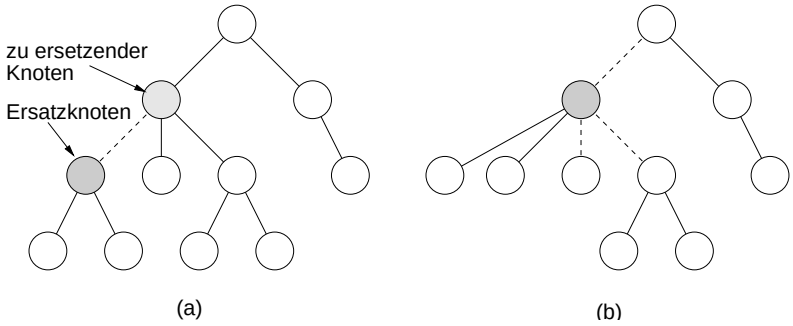


Abbildung 6.22.: Ersatz eines Knotens durch einen Sohnknoten: (a) Baum vor dem Ersatz, (b) Baum nach dem Ersatz.

Die Probleme eins und zwei werden zusammen in Unterkapitel 6.3.1 behandelt, wo ein Verfahren für den Ersatz von MCS während des laufenden Betriebes beschrieben wird. Das unter Punkt drei beschriebene Problem, dass nur ein Baum für eine Gruppe existiert, wird in Unterkapitel 6.3.2 behandelt. Es wird eine Lösung vorgestellt, die es erlaubt, mehrere Bäume innerhalb einer Gruppe einzusetzen.

6.3.1. Ersetzung eines primären MCS

In diesem Unterkapitel wird die Ersetzung eines MCS durch einen anderen MCS erläutert. Gründe für eine Ersetzung eines MCS können eine zu hohe Belastung oder ein Austritt aller Teilnehmer in einem lokalen Netz sein. Bevor auf die Gründe genauer eingegangen wird, soll das Grundprinzip der Ersetzung erklärt werden.

Ein Knoten im MCS-Baum wird ersetzt, indem einer der Sohnknoten seinen Platz einnimmt. In Abbildung 6.22 ist das prinzipielle Vorgehen dargestellt. Die gestrichelten Linien sind Verbindungen, die gelöscht (in Abbildung 6.22(a)) oder aufgebaut (in Abbildung 6.22(b)) werden müssen.

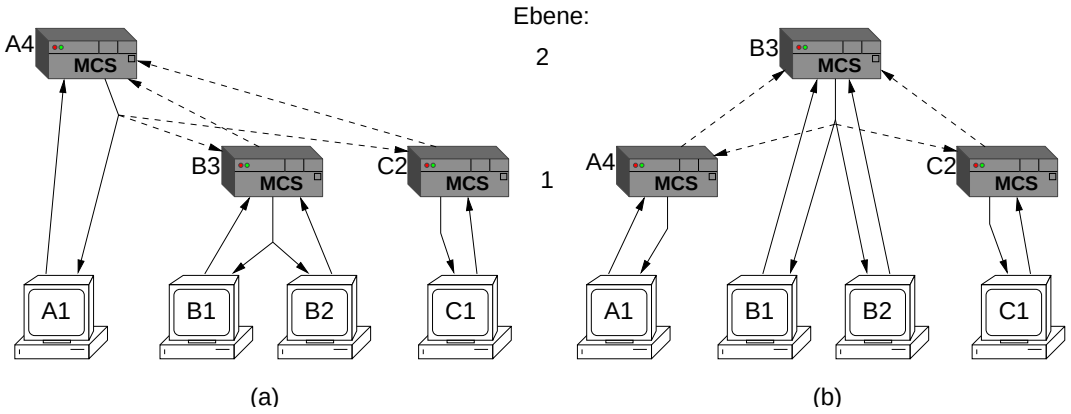


Abbildung 6.23.: Ersatz eines primären MCS: (a) Vor dem Ersatz von MCS A4 durch MCS B3 und (b) nach dem Ersatz.

Die Ersetzung eines MCS soll an einem Beispiel konkreter verdeutlicht werden. Das Beispiel in Abbildung 6.22 ist abstrakt, da dort nicht zwischen den verschiedenen ausgerichteten Verbindungen zwischen den Knoten unterschieden wird. Die Abbildung 6.23(a) zeigt einen zweistufigen Baum zur Gruppenkommunikation mit drei MCS A4, B3 und C2. Der MCS A4 ist primärer MCS der Ebene 2 und die MCS B3 und C2 sind primäre MCS der Ebene 1. Der MCS A4 soll durch den MCS B3 ersetzt werden. Wird der MCS A4 einfach entfernt, wäre die Gruppenkommunikation unterbrochen, da alle Daten über MCS A4 verteilt werden. Daher ist es notwendig, vor dem Entfernen von MCS A4 einen Ersatz als primären MCS der Ebene 2 zu bestimmen.

Als Ersatz kommen im Beispiel die beiden MCS B3 und C2 in Betracht. Wird beispielsweise der MCS B3 gewählt, dann müssen die gestrichelt dargestellten Verbindungen in Abbildung 6.23(a) abgebaut und die in Abbildung 6.23(b) ebenfalls gestrichelten Verbindungen neu aufgebaut werden.

In der Abbildung 6.23(b) ist als Ursache eine zu hohe Belastung des MCS A4 angenommen worden. Daher ist der MCS A4 nach der Ersetzung noch an der Gruppenkommunikation beteiligt. Allerdings ist er nur noch auf Ebene 1 primärer MCS, und MCS B3 ist neuer primärer MCS der Ebene 2. Ebenso könnte der MCS A4 aus der Gruppe austreten, wenn die Ursache der Austritt aller Teilnehmer ist. Es würden dann keine neuen Verbindungen zu MCS A4 aufgebaut werden, wodurch MCS A4 nicht mehr an der Gruppenkommunikation teilnimmt. Welcher Fall wann verwendet wird, hängt vom Ziel der Ersetzung ab.

Ist das Ziel der Ersetzung, den MCS aus der Gruppenkommunikation zu entfernen, so werden die Verbindungen zu MCS A4 nur abgebaut und keine neuen von oder zu MCS A4 aufgebaut. Dieses Ziel wird bei der Ersetzung verfolgt, wenn im lokalen Netz des MCS A4 keine lokalen Teilnehmer mehr vorhanden sind. Wird hingegen eine Ersetzung durchgeführt, um die Belastung in MCS A4 zu senken, so wird der MCS A4 nur als primärer MCS auf der Ebene 2 ausgewechselt.

Im Folgenden wird das Vorgehen bei der Ersetzung genauer erläutert. Zuerst wird die Reihenfolge bei der Ersetzung erklärt, wenn die Hierarchie zur Gruppenkommunikation aus mehr als zwei Ebenen besteht. Anschließend werden die einzelnen Schritte für eine erfolgreiche Ersetzung beschrieben. Doch zunächst soll das generelle Vorgehen beim Auf- und Abbau von ATM-Verbindungen an dieser Stelle beschrieben werden.

Auf- und Abbau von ATM-Verbindungen

Bei der Änderung der ATM-Verbindungen während die Gruppe aktiv ist stellt sich das Problem, ob zuerst die alten Verbindungen abgebaut und anschließend die neuen Verbindungen aufgebaut werden sollen oder umgekehrt. Wenn die alten Verbindungen zuerst abgebaut und dann die neuen Verbindungen aufgebaut werden, entsteht eine Unterbrechung im Datenstrom, wie in Abbildung 6.24(a) dargestellt. Im anderen Fall, wenn die neuen Verbindungen erst aufgebaut und danach die alten Verbindungen abgebaut werden können duplizierte Pakete entstehen (Abbildung 6.24(b)) und darüber hinaus besteht eine erhöhte Wahrscheinlichkeit, dass die neue Verbindung evtl. nicht aufgebaut werden kann, weil nicht genügend Ressourcen im ATM-Netz zur Verfügung stehen.

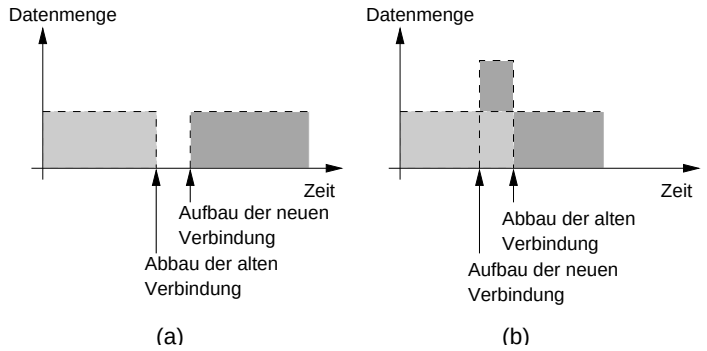


Abbildung 6.24.: Auf- und Abbau von ATM-Verbindungen: (a) Unterbrechung im Datenstrom und (b) Duplizierung von Daten.

Im weiteren Verlauf dieses Unterkapitels wird zuerst immer die alte Verbindung abgebaut und anschließend die neue Verbindung aufgebaut. Mit der Folge, dass der Datentransfer innerhalb einer Gruppe kurzzeitig unterbrochen wird (Abbildung 6.24(a)). Dieses Vorgehen ist gewählt worden, da es die einfachste Variante für die Realisierung darstellt. Das ist allerdings keine Einschränkung oder Nachteil bei dem hier vorgestellten Verfahren zur Ersetzung eines MCS. Es hat nur Einfluss auf die Reihenfolge der Aktionen, die bei der Ersetzung eines MCS ausgeführt werden müssen.

Reihenfolge bei der Ersetzung

Soll ein MCS aus dem Baum entfernt werden, so ist als erstes zu prüfen, ob er höherer primärer MCS ist und falls ja, für welche Ebenen. Ist er kein höherer primärer MCS, kann der MCS ohne Ersetzung entfernt werden. Hierzu reichen die in Unterkapitel 6.2.2 vorgestellten Aktualisierungen zwischen den Controllern aus.

Ist der zu entfernende MCS ein höherer primärer MCS, so müssen eine oder mehrere Ersetzungen durchgeführt werden. Mehrere Ersetzungen werden durchgeführt, wenn der zu ersetzende MCS primärer MCS der Ebene 3 und höher ist. Die Anzahl der Ersetzungen entspricht der Anzahl der Hierarchieebenen minus eins.

Wird mehr als eine Ersetzung durchgeführt, ist die Frage, in welcher Reihenfolge die Ersetzungen durchgeführt werden sollen. Hier kommt nur die Ersetzung beginnend mit der höchsten Hierarchieebene in Betracht. Eine Ersetzung beginnend mit der niedrigsten Hierarchieebene oder auf einer anderen Ebene als der höchsten ist nicht möglich. Da nach dem Prinzip der Abstraktion der Netzsegmente nur primäre MCS dem nächst höheren Controller bekannt sind. D. h. ist ein MCS auf einer Ebene x nicht primärer MCS, so ist er dem Controller der Ebene x+1 nicht bekannt und kann folglich auch nicht als der primäre MCS der Ebene x+1 ausgewählt werden.

Ablauf bei der MCS-Ersetzung

Die Ersetzung eines MCS läuft nach einem rekursiven Schema ab, welches in mehrere Schritte unterteilt wird. Die Schritte sollen hier im Überblick dargestellt und im folgen-

den Abschnitt genauer erläutert werden.

Auslöser für eine Ersetzung können drei verschiedene Ereignisse sein:

1. Der Austritt des letzten Teilnehmers in einem lokalen Netzwerk, in dem ein primärer MCS mindestens der Ebene 2 vorhanden ist, löst eine Ersetzung auf allen Ebenen größer 1 bis hin zur höchsten Hierarchieebene aus, für die der MCS verantwortlich ist.

2. Eine zu hohe Belastung in einem primären MCS der Ebene 2 und größer kann zu einer Ersetzung auf einer oder mehreren höheren Ebenen führen. Ein primärer MCS auf Ebene 1 kann nciht ersetzt werden, da dieser die lokalen Teilnehmer bedienen muss.

3. Für einen MCS, der in mehreren Bäumen primärer MCS ist und mindestens in einem davon auf Ebene 2, kann eine Ersetzung in einem Baum auf Ebene 2 oder höher ausgelöst werden. Dieser Fall wird erst in Unterkapitel 6.3.2 beschrieben, wenn eine Gruppenkommunikation basierend auf mehreren Bäumen vorgestellt wird.

Bei dem zweiten Ereignis ist der Auslöser eine zu hohe Belastung (siehe auch Unterkapitel 5.3.3 ab Seite 70) in einem primären MCS. Als Kriterium für eine zu hohe Belastung dient ein vorher festgelegter Schwellwert, bei dessen Überschreitung eine MCS-Ersetzung eingeleitet wird. Die Wahl des Schwellwertes hat keinen Einfluss auf das eigentliche Verfahren zur MCS-Ersetzung.

Tritt eines dieser drei Ereignisse ein, wird bzw. kann eine Ersetzung ausgelöst werden. Die Ersetzung erfolgt in den Schritten wie sie Abbildung 6.25 zeigt. Im ersten Schritt sendet der Controller, in dem das Ereignis auftrat, eine Replace-Aufforderung (vgl. Anhang C ab Seite 177) an den nächst höheren Controller. Die Replace-Aufforderung wird solange in der Hierarchie nach oben weitergeleitet, bis der Controller gefunden ist, in dem der zu ersetzende MCS der primäre MCS ist, aber nicht mehr primärer MCS eines weiteren, höheren Controllers.

Hat der zuständige Controller die Aufforderung zur Ersetzung erhalten, müssen zunächst die möglichen Ersatz-MCS ermittelt werden (Schritt 3). Wenn der Ersatz-MCS bestimmt ist, können die Verbindungen entsprechend ab- und aufgebaut werden (Schritt 4). Sind die Verbindungen korrigiert, muss überprüft werden, ob weitere Ersetzungen notwendig sind (Schritt 5).

Für den Fall, dass weitere Ersetzungen notwendig sind, wird eine neue Replace-Aufforderung an den nächst niederen Controller gesendet (Schritt 6) und eine neue Ersetzung beginnt bei Schritt 2. Sind keine weiteren Ersetzungen notwendig, so ist die Ersetzung eines MCS abgeschlossen.

Auswahl eines Ersatz-MCS

Bevor ein MCS ersetzt werden kann, muss hierfür ein Ersatz-MCS ermittelt werden (Schritt 3 in Abbildung 6.25). Als Ersatz-MCS für einen primären MCS der Ebene

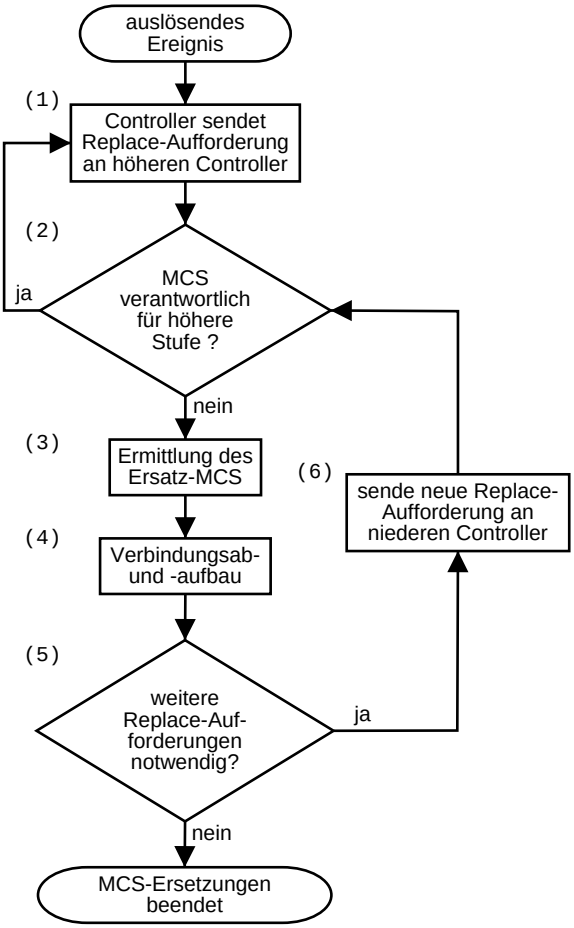


Abbildung 6.25.: Ablauf der MCS-Ersetzung.

x+1 kommen alle primären MCS der Ebene x in Betracht. Bedingt ist dies durch das Abstraktionskonzept, da ein primärer MCS der Ebene x auch immer primärer MCS aller Ebenen kleiner x ist.

Konkret können die Ersatz-MCS folgendermaßen ausgewählt werden: Der zu ersetzende MCS in Ebene x+1 ist im zuständigen Controller der primäre MCS in der Detailschicht und es kommen als Ersatz-MCS alle (nicht primären) MCS der Detailschicht in Frage, die nicht in der Aggregationsschicht enthalten sind. Sind sie in der Aggregationsschicht enthalten, bedeutet das, dass sie bereits primäre MCS in einem anderen Netzsegment sind, welche über den höheren Controller angesprochen werden können, und kommen somit nicht in Betracht.

Sind mehrere Ersatz-MCS ermittelt worden, muss einer von diesen ausgewählt werden. Hierzu könnte einer der MCS wahllos genommen oder anhand seiner aktuellen Belastung bestimmt werden. Um die Anzahl der zukünftigen Ersetzungen gering zu halten, scheint die Auswahl anhand der Belastung die bessere, wenn auch etwas aufwändigere, Methode zu sein. Um die Belastung zu ermitteln, müssen zunächst Anfragen an alle Ersatz-MCS gesendet werden. Die Belastungsabfrage eines MCS erfolgt wie beim

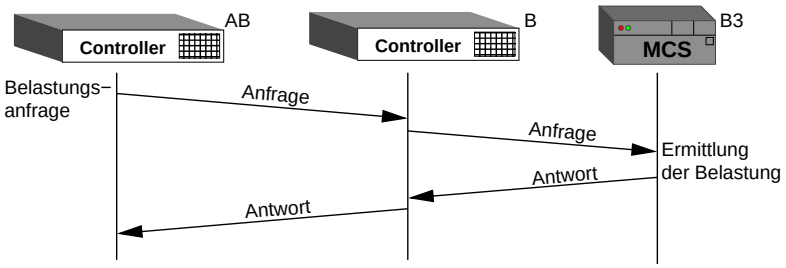


Abbildung 6.26.: Schema der Belastungsanfragen und -antworten.

lokalen MCS-Schema (Unterkapitel 5.4.2 ab Seite 74), nur dass jetzt die Abfrage über mehrere Controller bis zum jeweiligen MCS weitergeleitet werden muss. Das ist in Abbildung 6.26 als Weg-Zeit-Diagramm dargestellt. Sind alle Antworten von den MCS beim Controller eingetroffen, wird der neue Ersatz-MCS ausgewählt.

Verbindungsaufbau und -abbau

Der eigentliche Prozess der Ersetzung ist der Ab- und Aufbau der ATM-Verbindungen (Schritt 4 in Diagramm 6.25) beim Ersetzen des primären MCS. Bei dem Ersetzen eines primären MCS auf einer Hierarchieebene sind die darüber liegenden und die darunter liegenden Hierarchieebenen beteiligt, wobei die darüber liegende Ebene nicht immer vorhanden sein muss. Das ist z. B. in dem Szenario in Abbildung 6.23 der Fall.

Um den Wechsel der Verbindungen zu veranlassen, müssen mehrere Controller aktualisiert werden. Zuerst muss der Controller aktualisiert werden, in dem die Ersetzung stattfindet. Ist dieser Controller aktualisiert, müssen alle benachbarten Controller in den höheren und unteren Ebenen aktualisiert werden. Durch diese Aktualisierungen werden zuerst die nicht mehr benötigten ATM-Verbindungen abgebaut und anschließend die neuen Verbindungen aufgebaut. Der Abbau und Aufbau der ATM-Verbindungen erfolgt in den beteiligten MCS nicht immer zeitgleich. Das ist bedingt durch unterschiedliche Verzögerungen auf den Signalisierungsverbindungen. Diese zeitlichen Verschiebungen können dazu führen, dass eine Unterbrechung in der Gruppenkommunikation entsteht, von der alle Empfänger betroffen sind.

6.3.2. Lastverteilung durch parallele Bäume

Das zentrale Ziel beim sender-orientierten Schema für lokale ATM-Netze ist es, die Belastung von einem MCS auf mehrere zu verteilen (vgl. Kapitel 5), um so die Datenkonzentration in einem MCS reduzieren zu können. Dieser Vorteil des sender-orientierten Schemas ging bei der Integration in die hierarchische Gruppenkommunikation verloren. Bedingt ist das dadurch, dass auf jeder höheren Hierarchieebene immer nur ein einziger MCS pro Gruppe der primärer MCS sein kann. Eine Belastungsverteilung auf mehrere MCS innerhalb einer Gruppe ist auf den höheren Hierarchieebenen daher nicht möglich.

Die Grundidee bei der Einführung von parallelen Bäumen (in der Graphentheorie auch als Wald bezeichnet) zeigt Abbildung 6.27. Ein neuer Baum entsteht immer durch

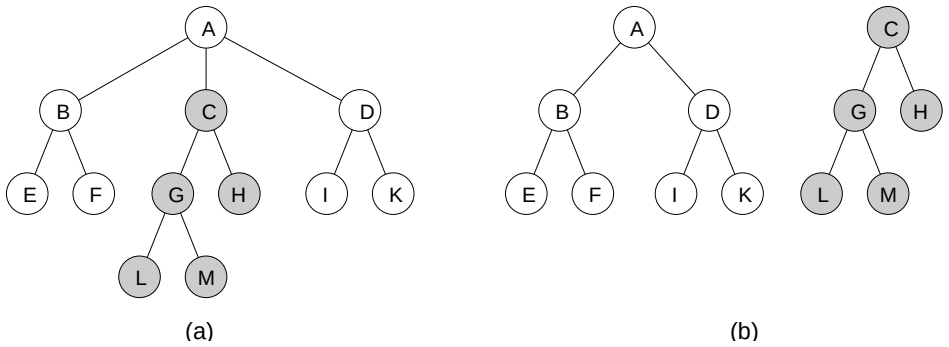


Abbildung 6.27.: Aufspaltung eines Baumes: (a) ursprünglicher Baum und (b) zwei neue Bäume.

Aufspaltung eines bereits bestehenden Baumes in zwei neue Bäume. In der Abbildung 6.27 ist hierzu ein kompletter Teilbaum abgespalten worden, was die einfachste, aber nicht immer idealste, Vorgehensweise für die Aufteilung eines Baumes darstellt. Darüber hinaus kann sich die Höhe des ursprünglichen Baumes verringern, was bedeutet, dass weniger Zwischensysteme beteiligt und somit die Verzögerung verringert werden kann.

Da die Aufspaltung eines Baumes in zwei neue Bäume das Datenaufkommen pro Baum reduziert (im Idealfall um 50%), wird somit auch die Belastung und damit einhergehend die Verzögerung in den MCS gemindert. Ein anderer Vorteil ergibt sich durch den Einsatz mehrerer lokaler MCS in mehreren Bäumen. Bisher konnte nur ein lokaler MCS pro Baum eingesetzt werden. Diese Einschränkung bleibt bestehen, aber bei mehreren Bäumen können jetzt jedem Baum lokal unterschiedliche MCS zugeordnet werden. Unabhängig von den MCS muss aber immer die Bedingung erfüllt sein, dass die Bäume in den Endsystemen zusammenlaufen. Nur dann erhält jeder Empfänger auch alle Daten der Gruppe.

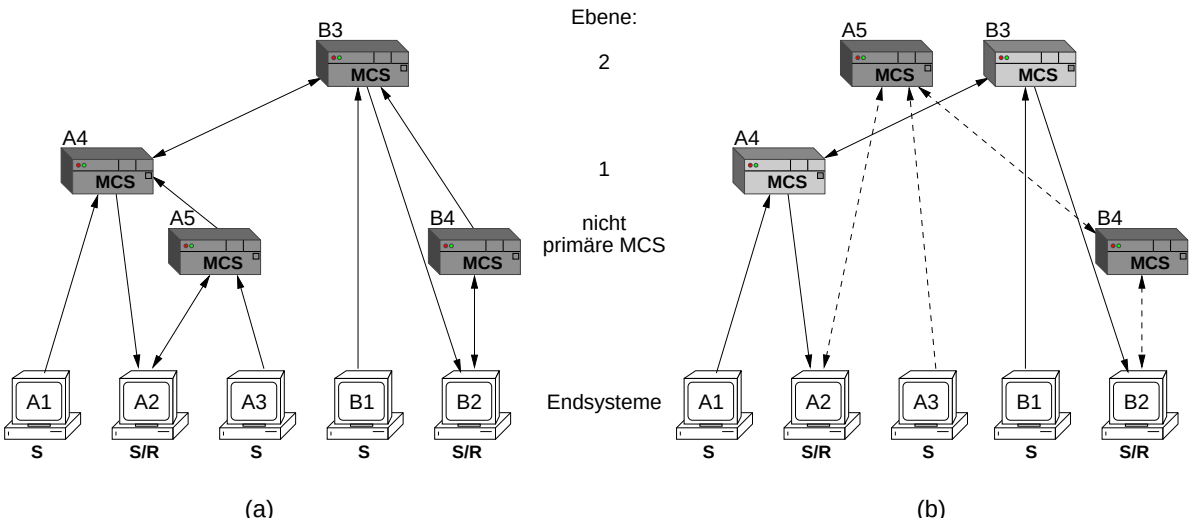


Abbildung 6.28.: Gruppenkommunikation (a) mit einem Baum und (b) mit zwei Bäumen.

Ein Beispiel soll dies verdeutlichen. Abbildung 6.28(a) zeigt einen einzelnen Baum zur Gruppenkommunikation. Das Gruppenkommunikationsschema wird in diesem Beispiel nur hierarchisch dargestellt. Die Punkt-zu-Punkt- und Punkt-zu-Mehrpunkt-Verbindungen werden nicht mehr gezeigt, damit die Übersichtlichkeit erhalten bleibt. Stattdessen wird eine Kommunikation in beiden Richtungen durch eine Verbindung mit Pfeilen an beiden Enden symbolisiert. Die Hierarchie besteht aus zwei Ebenen und zwei lokalen Netzen. Der höchste primäre MCS ist B3 auf Ebene 2. Daneben gibt es den primären MCS A4 auf Ebene 1. Die MCS A5 und B4 sind lokale, nicht primäre MCS.

Abbildung 6.28(b) zeigt für die gleiche Hierarchie das Gruppenkommunikationsszenario mit zwei parallelen Bäumen. Die Verbindungen des zweiten Baums sind durch gestrichelte Verbindungen dargestellt. Idealerweise sind in diesem Szenario in jedem lokalen Netz genau zwei MCS vorhanden. Das macht es besonders einfach, den zweiten Baum aufzubauen. Steht nicht in jedem lokalen Netz für jeden Baum ein expliziter MCS zur Verfügung, so muss ein MCS bestimmt werden, der mehreren Bäumen angehört. Die Problematik, die hierdurch entsteht, wird weiter unten in diesem Abschnitt beschrieben.

Im Beispiel in Abbildung 6.28(b) sind die MCS A5 und B3 höhere primäre MCS der Ebene 2, und die MCS A4 und B4 sind primäre MCS der Ebene 1. Dabei gehören die MCS A4, B3 zum ersten und die MCS A5, B4 zum zweiten Baum. Ferner sind die Sender so verteilt, dass jedem MCS mindestens ein Sender zugeteilt worden ist. Die gleichmäßige Verteilung der Sender auf die MCS ist besonders wichtig, damit die Daten optimal auf alle Bäume verteilt werden können.

Im Gegensatz zur prinzipiellen Darstellung der Aufteilung in Abbildung 6.27 ist in Abbildung 6.28(b) zu erkennen, dass die Bäume in den Endsystemen zusammenlaufen. Das ist auch unmittelbar notwendig, da ansonsten nicht alle Empfänger dieselben Daten erhalten können. Die Endsysteme stellen aber nicht die einzig möglichen Berührungspunkte zwischen zwei Bäumen dar. Sind zwei Bäume vorhanden, in einem lokalen Netz aber nur ein einziger MCS, so muss dieser MCS in beiden Bäumen aktiv sein. Dieser Fall wird weiter unten in diesem Kapitel behandelt (ab Seite 125). Grundsätzlich gilt, dass dieser Fall nach Möglichkeit vermieden werden sollte, da ansonsten die oben beschriebenen Vorteile von parallelen Bäumen nicht greifen können.

Bisher wurde nur erwähnt, dass der Aufbau eines weiteren Baums dynamisch erfolgen soll. Dynamisch heißt, dass ein weiterer Baum während der laufenden Kommunikation einer Gruppe bei Bedarf aufgebaut wird. Der Bedarfsfall tritt ein, wenn die Belastung einzelner MCS vordefinierte Grenzen (z. B. ein durch technische Realisierung begrenztes Belastungsmaximum) überschreitet und ausreichend viele MCS im ATM-Netz für diese Gruppe zur Verfügung stehen. D. h. bei Entstehung einer Gruppe ist beispielsweise der in Abbildung 6.28(a) dargestellte Baum vorhanden. Durch eine erhöhte Belastung in einzelnen MCS wird der Aufbau eines neuen Baums initiiert und das in Abbildung 6.28(b) dargestellte Schema zur Gruppenkommunikation entsteht.

Im Folgenden soll der dynamische Aufbau eines weiteren Baums zuerst grob und anschließend in allen Schritten im Detail erläutert werden. Abschließend wird noch auf den Abbau eines Baums eingegangen.

Aufbau eines neuen Baums

Beim Aufbau eines neuen Baums müssen eine Reihe von Bedingungen und Aktionen ausgeführt werden:

1. In einem lokalen Netz wird festgestellt, dass die Belastung eines MCS zu hoch ist. Das initiiert den Aufbau eines neuen Baums. Der höchste Controller im Baum wird von der lokal überhöhten MCS-Belastung informiert.

2. Im höchsten Controller des Baums wird in allen Unterbäumen geprüft, ob ausreichend unbelastete MCS für einen weiteren Baum zur Verfügung stehen.

3. Sind ausreichend MCS vorhanden, so wird der Aufbau eines neuen Baums eingeleitet.

Im Detail unterteilt sich der Aufbau eines neuen Baums in fünf Schritte. Den gesamten Ablauf des Aufbaus zeigt das Flussdiagramm in Abbildung 6.29.

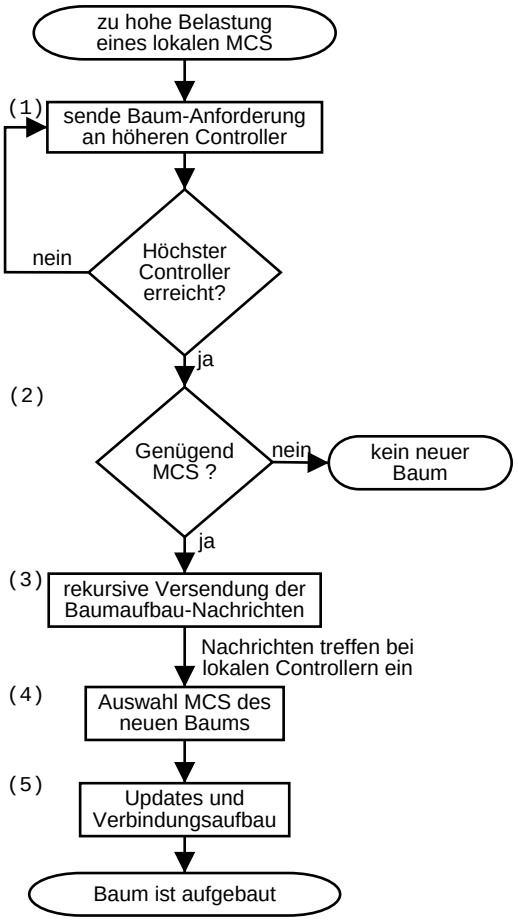


Abbildung 6.29.: Ablauf des Aufbaus eines neuen Baums.

Der Auslöser des Aufbaus ist eine zu hohe Belastung eines oder mehrerer MCS. Eine zu hohe Belastung wird registriert, indem die Belastung eines MCS einen vorgegebenen

Schwellwert überschreitet. Der Schwellwert ist wie bei der MCS-Ersetzung vorgegeben, hat aber keinen Einfluss auf das generelle Vorgehen zum Aufbau eines neuen Baums.

Ist die Belastung eines MCS zu hoch, wird vom lokalen Controller eine Aufforderung einen neuen Baum zu bilden an den nächst höheren Controller gesendet (Schritt 1). Dieser Vorgang wird wiederholt, bis der höchste Controller erreicht ist. Beim höchsten Controller angekommen, wird geprüft, ob genügend MCS zur Verfügung stehen, um einen neuen Baum aufzubauen (Schritt 2).

Ist auch diese Bedingung erfüllt, werden die Baumaufbau-Nachrichten an alle lokalen Controller gesendet (Schritt 3). Dort wird ein geeigneter MCS ausgewählt, der dem neuen Baum angehören soll (Schritt 4) und im lokalen Controller eingetragen. Im fünften und letzten Schritt werden die Rückmeldungen versendet und der Baum bzw. die Verbindungen etabliert.

Nachdem der Gesamtablauf grob erläutert wurde, werden der Auslöser und die einzelnen Schritte im Detail behandelt:

Belastungsmitteilung vom MCS Um aktuelle Belastungen von den MCS zu erhalten, versendet jeder MCS Nachrichten über seine momentane Belastung an den lokalen Controller. Da die Belastungen im MCS periodisch gemessen werden, wird nur eine Belastungsnachricht an den lokalen Controller gesendet, wenn die Änderung zur letzten gesendeten Belastung eine Grenze von $\pm 10\%$ überschreitet. Die Formel lautet (B_{new} = aktuelle Belastung, B_{old} = letzte gemeldete Belastung):

$$\frac{|B_{old} - B_{new}|}{B_{old}} > 0.1, B_{old} > 0$$

Der lokale Controller wertet die aktuelle Belastung des MCS aus und sendet eine Belastungsnachricht (vgl. Anhang C ab Seite 177) an den höheren Controller, wenn die aktuelle Belastung den vorgegebenen Schwellwert überschreitet.

Anfrage nach neuem Baum Übersteigt die Belastung eines MCS ein definiertes Maximum (s.o.), wird eine Aktualisierung an den nächst höheren Controller in der Hierarchie gesendet. Mit dieser Aktualisierung wird zusätzlich die Anzahl der lokal vorhandenen MCS gesendet.

Trifft die Nachricht beim nächst höheren Controller ein, speichert dieser ebenfalls die erhaltenen Daten. Zusätzlich wird die Anzahl der MCS und der lokalen Netze, die unterstützt werden, gespeichert. Diese Informationen werden zusammen mit den Belastungen der MCS an den nächst höheren Controller gesendet. Dieses Weitersenden erfolgt so lange, bis die Nachricht beim höchsten Controller eintrifft.

Bedingungsprüfung im höchsten Controller Empfängt der höchste Controller im Baum eine Nachricht über den Aufbau eines neuen Baumes, muss eine Überprüfung durchgeführt werden, ob ausreichend MCS für die Etablierung eines zusätzlichen Baumes zur Verfügung stehen (siehe hierzu auch Seite 126 'Weniger lokale MCS als Bäume'). Hierzu wird eine Anfrage an alle untergeordneten Controller gesendet, um die aktuelle Anzahl der MCS und aktiven lokalen Netze zu erfahren. Dieses

Vorgehen ist notwendig, da nur die Information aus dem Teilbaum mit dem überlasteten MCS vorliegt, aber Informationen aus allen Teilbäumen benötigt werden.

Das Versenden erfolgt analog zu dem in Abbildung 6.26 auf Seite 119 beschrieben Verfahren. Sind alle Antworten mit den aktuellen Daten der MCS und lokalen Netze eingetroffen, so kann ermittelt werden, ob genügend MCS pro lokalem Netz zur Verfügung stehen. Die Formel hierzu ist:

$$z = \frac{m}{s * (t + 1)}$$

Hierbei bedeutet m die Anzahl der MCS, s die Anzahl der lokalen Netze und t ist die Anzahl vorhandener Bäume. Das Ergebnis z ist also die Anzahl der MCS im Verhältnis zur gewünschten Anzahl MCS in allen lokalen Netzen. Die gewünschte Anzahl MCS ergibt sich aus dem Ideal, dass für jeden Baum ein separater MCS zur Verfügung stehen sollte. Ist das Verhältnis z größer gleich eins ($z \geq 1$), so wird ein neuer Baum erzeugt. Bei $z < 1$ ist es möglich einen Baum bei einer besonders hohen Belastung einiger MCS zu erzeugen. Bei einem zu kleinen Wert von z wird kein neuer Baum erzeugt, da dann nicht genügend MCS für einen weiteren Baum vorhanden sind.

Der Nachteil dieses Verfahrens ist, dass nicht ermittelt werden kann, wie die MCS auf die lokalen Netze verteilt sind. Im obigen Szenario (Abbildung 6.28(a)) sind z. B. jeweils zwei MCS pro lokalem Netz vorhanden, welches die ideale Voraussetzung für zwei Bäume darstellt. Ebenso können aber im obigen Szenario drei MCS im lokalen Netz A vorhanden sein und nur einer im Netz B oder umgekehrt. In allen drei Fällen würde der Wert z gleich 1 sein. Aufgrund der Abstraktion der Netzsegmente ist es aber nicht möglich, alle lokalen Netze einzeln aufzuführen. Stattdessen ist es nur möglich, die lokalen Netze aggregiert darzustellen. Hierdurch kann aber gerade eine ungleichmäßige Verteilung nicht erfasst werden.

Ein weiter Nachteil besteht in der zeitlichen Verzögerung, die durch die Anfragen entsteht. In diesem Zeitraum kann sich die Belastung wieder reduziert haben. In diesem Fall wird aber dennoch ein neuer Baum aufgebaut. Da die gemessenen Belastungswerte in den MCS zeitlich gewichtet werden (Unterkapitel 5.3.3, Seite 70), kann aber in der Regel von einer länger vorausgehenden Belastung eines oder mehrerer MCS ausgegangen werden, die mit größerer Wahrscheinlichkeit auch noch anhalten wird.

Baumaufbau-Nachrichten versenden Ist die Bedingungsprüfung im höchsten Controller positiv ausgefallen, müssen die Anweisungen für den Aufbau des Baums versendet werden. Ein neuer Baum wird immer ausgehend von der niedrigsten Hierarchieebene zur höchsten aufgebaut. Um diesen Vorgang einzuleiten, müssen Nachrichten vom höchsten Controller an alle lokalen Controller gesendet werden. Die versendete Nachricht entspricht einer Anweisung, die den lokalen Controller auffordert, einen lokalen MCS zu bestimmen, der dem neuen Baum angehören soll.

Die Weiterleitung der Nachrichten erfolgt wieder nach dem in Abbildung 6.26 auf Seite 119 vorgestellten Prinzip. Der daraufhin eingeleitete Vorgang des Aufbaus eines neuen Baumes erfolgt nach dem bekannten Vorgehen bei der Signalisierung aus Unterkapitel 6.2.2, Seite 100.

MCS-Auswahl für den neuen Baum Trifft in einem lokalen Controller die Baumaufbau-Nachricht ein, beginnt dieser Controller damit, einen lokalen MCS zu bestimmen, der in den neuen Baum integriert werden soll. Im Optimalfall ist ein MCS vorhanden, der nicht primärer MCS in diesem Controller ist (damit ist er auch automatisch in keinem anderen Controller ein primärer MCS). Sind mehrere nicht primäre MCS vorhanden, wird der mit der geringsten Belastung ausgewählt.

Sind aber nur primäre MCS vorhanden, muss einer der primären MCS der anderen Bäume gewählt werden. Hierbei wird ebenfalls der MCS mit der geringsten Belastung gewählt. Bei der Gruppenkommunikation in Abbildung 6.28(a) wird z. B. immer MCS A5 gewählt, da dieser kein primärer MCS ist.

Verbindungsab- und -aufbau

Der neue Baum besteht im Allgemeinen aus neu hinzugekommenen und bereits in der Gruppe aktiven MCS. Dabei sollte die Anzahl der bereits aktiven MCS in der Gruppe so gering wie möglich sein, was über die Bedingungsprüfung im höchsten Controller (s.o.) annähernd erreicht werden kann.

Bei den bereits aktiven MCS bauen diese nach Möglichkeit zuerst ihre alten Verbindungen ab, bevor die Verbindungen zum neuen Baum etabliert werden. Das ist aber nicht immer möglich, und zwar dann, wenn der gewählte MCS primärer MCS eines Baums der Gruppe ist. In diesem Fall können die Verbindungen zu dem oder den anderen Bäumen nicht abgebaut werden und der MCS ist in zwei Bäumen derselben Gruppe aktiv.

Beim Verbindungsab- und -aufbau muss zwischen verschiedenen Fällen unterschieden werden. Dabei kommt es auf die Anzahlen der MCS und etablierten Bäume an. Es wird zwischen gleich vielen MCS und Bäumen, mehr MCS als Bäume und weniger MCS als Bäume unterschieden:

Gleich viele lokale MCS und Bäume Sind in einem lokalen ATM-Netz genau so viele MCS vorhanden, wie die Anzahl der zu bildenden Bäume, gibt es keine Probleme. Zuerst werden alle Verbindungen des gewählten MCS zu den MCS des alten Baums abgebaut. Anschließend werden die neuen Verbindungen zu den anderen MCS des neuen Baums aufgebaut.

Im Zeitraum zwischen dem Verbindungsabbau und -aufbau entsteht wie bei der MCS-Ersetzung eine Unterbrechung in der Gruppenkommunikation.

Mehr lokale MCS als Bäume Sind mehr lokale MCS als Bäume vorhanden, gibt es ebenfalls keine Probleme. Jeweils ein MCS muss jedem Baum angehören. Die restlichen MCS können beliebig aufgeteilt werden. Der Verbindungsab- und -aufbau erfolgt nach den bisher vorgestellten Schemata.

Weniger lokale MCS als Bäume Wenn weniger lokale MCS als Bäume vorhanden sind, besteht das Problem, dass mindestens ein MCS mehreren Bäumen angehören muss. Damit ergibt sich ein Berührungspunkt der Bäume vor den Empfängern. Dieser Fall kann nicht immer vermieden werden.

Das Problem besteht darin, dass der MCS die Verbindungen beider Bäume getrennt voneinander zu behandeln hat. D. h. es dürfen beispielsweise nur Daten des Baums 1 auf Verbindungen des Baums 1 gemultiplext werden und nicht auf Verbindungen anderer Bäume, da ansonsten die Daten dupliziert werden. Hierzu muss der MCS intern die Bäume separat voneinander behandeln und auch für jeden Baum getrennte Verbindungen verwalten.

Primäre MCS in mehreren Bäumen Wenn nicht für jeden Baum lokal ein exklusiver MCS zur Verfügung steht, können MCS mit mehreren Bäumen nicht vermieden werden. Das hat weiterhin zur Folge, dass es auch primäre MCS gibt, die für mehrere Bäume verantwortlich sind. Diese MCS weisen eine erhöhte Belastung auf, da sie außer den Daten von den lokalen Sendern auch die Daten untergeordneter MCS weiterleiten müssen. Das hat zur Folge, dass für jede weitere Ebene, für die ein MCS in mehreren Bäumen primärer MCS ist, die Belastung dieses MCS steigt. Daher sollten nach Möglichkeit primäre MCS für mehrere Bäume vermieden werden.

Ist die Belastung eines solchen primären MCS zu hoch, ist es zwecklos, diesen MCS durch Aufbau eines weiteren neuen Baums entlasten zu wollen, da höchstwahrscheinlich der betroffene MCS auch am neuen Baum partizipieren wird. Es besteht aber die Möglichkeit, den überlasteten MCS in einem Baum auszutauschen, wie bereits in Unterkapitel 6.3.1 beschrieben. Das dort für einen Baum pro Gruppe dargelegte Vorgehen ist ohne Änderungen auch für einen von mehreren Bäumen einer Gruppe anwendbar.

Abbau eines Baums

Bisher ist nur beschrieben worden, wie und wann ein neuer Baum aufgebaut wird. Der umgekehrte Vorgang, der Abbau eines Baums bzw. die Zusammenführung zweier Bäume, kann aber genauso notwendig sein. Der Einsatz mehrerer Bäume pro Gruppe ist sinnvoll, wenn viele aktive Sender in einer Gruppe existieren und ein hohes Datenaufkommen vorhanden ist. In so einem Fall können mit mehreren Bäumen die Daten dezentraler organisiert und zwischen den Gruppenteilnehmern verteilt werden. Auf der Gegenseite erfordert jeder zusätzliche Baum weitere logische und physikalische Ressourcen im ATM-Netz. Daher ist es ebenso sinnvoll, einen Baum wieder abzubauen, wenn das Datenaufkommen signifikant nachgelassen hat.

Eine andere Möglichkeit wäre, den Baum nicht explizit abzubauen, sondern einfach zu warten, bis der Baum keine Teilnehmer mehr hat und damit implizit abgebaut ist. Dieses Vorgehen würde die Implementierung vereinfachen, aber auf der anderen Seite zu einer unnötig hohen Reservierung von ATM-Verbindungen führen

Die Abbildung 6.30 zeigt den Auslöser und die fünf Schritte des Abbaus. Der Auslöser ist eine geringe Belastung eines lokalen MCS, was analog zu einer erhöhten Belastung, durch Unterschreitung eines Schwellwertes erreicht wird. Der Ablauf beim Abbau eines Baums erfolgt ganz ähnlich zum Ablauf beim Aufbau eines Baums.

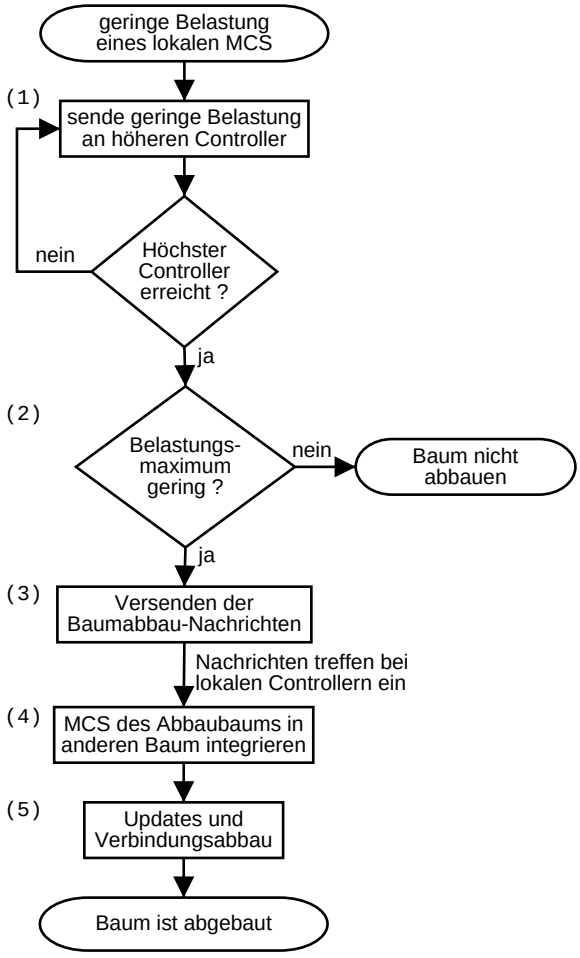


Abbildung 6.30.: Ablauf des Abbaus eines Baums.

Durch die Meldung einer geringen Belastung eines MCS wird vom lokalen Controller eine Abbauanfrage an den höheren Controller gesendet (Schritt 1). Diese Nachricht wird so lange weitergeleitet bis sie beim höchsten Controller angekommen ist. Dieser Controller prüft, ob das Belastungsmaximum wirklich gering ist (Schritt 2). Ist das der Fall, wird damit begonnen, Abbauanweisungen an alle lokalen Controller zu senden (Schritt 3). Die Weiterleitung der Anweisungen entspricht der Weiterleitung beim Aufbau des Baums. Treffen die Abbauanweisungen in den lokalen Controllern ein, so werden die noch aktiven MCS des zugehörigen Baums in andere Bäume integriert (Schritt 4). Das sind alle MCS, die noch Teilnehmer der Gruppe bedienen. Um das zu erreichen, kann es notwendig sein, eine MCS-Ersetzung (Unterkapitel 6.3.1) durchzuführen. Dieses ist genau dann der Fall, wenn der MCS ein primärer MCS einer höheren Ebene ist und noch andere zu unterstützende MCS im Baum vorhanden sind. Sind die MCS in die anderen

Bäume integriert, werden Aktualisierungen an die höheren Controller gesendet (Schritt 5). Erfolgen alle Aktualisierungen korrekt, ist der Baum abgebaut.

6.4. Leistungsbewertung

In diesem Unterkapitel werden die in den vorangegangenen Unterkapiteln vorgestellten Ansätze für eine globale Gruppenkommunikationsunterstützung von SkaGAN und das daraus entwickelte Modell auf dessen Leistungsfähigkeit untersucht. Für eine Bewertung werden dieselben Kriterien wie in Kapitel 3 (Stand der Forschung) herangezogen. Einige Ergebnisse sind daher auch schon in die Tabelle aus Unterkapitel 3.4 (ab Seite 51) mit eingegangen.

Zunächst soll erläutert werden, welche Messungen durchgeführt worden sind, und welchem Zweck die Messungen dienen sollen. Eine erste Einteilung der Messreihen ist durch die verschiedenen Teilgebiete (Gruppenverwaltung, Datentransfer und Erweiterungen) möglich. Zu den jeweiligen Teilgebieten werden die folgenden Messreihen durchgeführt:

Gruppenverwaltung: Das vorgestellte Schema der hierarchischen Gruppenverwaltung bei SkaGAN kann das zu verwaltende Datenaufkommen für eine Gruppe reduzieren. Es soll ermittelt werden, wie groß der Aufwand für die Verwaltung von Gruppen unterschiedlicher Größe und topologischer Verteilung ist.

Datentransfer: Das grundlegende Schema für den Datentransfer zwischen den Gruppenteilnehmern besteht aus einem gemeinsamen Baum, über den alle Empfänger die Daten erhalten. In diesem Zusammenhang interessieren vor allem die Anmeldeverzögerung, die Belastung der Zwischensysteme (MCS) und die Organisation des gemeinsamen Baumes im Netz.

Erweiterungen: Die Erweiterungen für eine verbesserte Lastverteilung ermöglichen durch Rekonfiguration des bestehenden Baumes, die Strukturen an die aktuelle Situationen anzupassen. Es soll ermittelt werden, inwieweit diese Anpassung erfolgt und welche Vor- und Nachteile (z. B. Unterbrechungen in der Kommunikation) für den Datentransport daraus entstehen.

Das Ziel der oben vorgestellten Messreihen ist es, die Funktionsfähigkeit des hier vorgestellten Ansatzes zur Gruppenkommunikation über ATM zu demonstrieren und die inhärenten Eigenschaften des Ansatzes festzustellen. Eine ganze Reihe weiterer möglicher Messungen, die von zusätzlichen Parametern (z. B. Link-Bandbreite und Hintergrundverkehr) abhängig sind, sind daher unterlassen worden. Das wären unter anderem:

Durchsatzrate: Auf dieses Kriterium ist verzichtet worden, da die maximale Durchsatzrate nur zum Teil von dem hier vorgestellten Ansatz begrenzt wird. Weitere begrenzende Faktoren sind die Leitungskapazitäten (inklusive Datenverkehr im Hintergrund) und die Belastbarkeit der Endsysteme. Insbesondere bei heterogenen Teilnehmern ist die Bewertung der Durchsatzrate zweifelhaft, da hier nicht

immer das Datenvolumen ausschlaggebend ist. Wichtiger für die Durchsatzrate ist die Kommunikationsstruktur, also inwieweit Datenkonzentrationen stattfinden und der Verteilbaum die kürzesten Wege zwischen den Teilnehmern ermöglicht.

Verzögerung und Verzögerungsschwankungen: Bei einer Gruppe variieren die Entfernungen der Teilnehmer untereinander. Dies macht sich vor allem durch eine unterschiedliche Anzahl von Vermittlungssystemen zwischen den Teilnehmern bemerkbar. Daher unterscheiden sich die Verzögerungen bei den Empfängern, je nachdem wieweit diese von den Sendern entfernt sind. In gleichen Maße können auch unterschiedliche Verzögerungsschwankungen auftreten. Es erscheint daher sinnvoller, ähnlich wie bei der Durchsatzrate, die Kommunikationsstruktur zu bewerten, da diese der wichtigste Einflussfaktor auf die Verzögerung ist.

6.4.1. Bewertung der Gruppenverwaltung

Die in Unterkapitel 6.1 beschriebene Gruppenverwaltung basiert auf einer Baumstruktur, wobei in jedem Baumknoten die Gruppenteilnehmer der Unterbäume aggregiert werden. Damit soll eine Reduktion des zu verwaltenden Datenaufkommens erreicht werden, wodurch das System auch für große Gruppen skalieren kann.

Um die Leistungsfähigkeit dieses Ansatzes beurteilen zu können, sind eine Reihe von Messungen am Modell durchgeführt worden. Um die wesentlichen Eigenschaften hervorzuheben sind hierzu einige Vereinfachungen vorgenommen worden:

- Die Gruppenverwaltung unterscheidet zwischen Sendern und Empfängern einer Gruppe. Das führt aus Sicht der Verwaltung letztendlich zu einer Separierung einer Gruppe in zwei Untergruppen, eine für Sender und eine für Empfänger. Für die folgenden Messungen ist nur eine Untergruppe (Sender oder Empfänger) untersucht worden. Die Ergebnisse müssen dann nur auf beide Untergruppe übertragen werden, die i. Allg. aber in einer Gruppe in unterschiedlichen Anteilen vorhanden sind.

- Es wird nur die Anmeldung von Gruppenteilnehmern gemessen. Die Teilnehmer können für die Messungen nur einer Gruppe beitreten und diese anschließend nicht mehr verlassen. Die Signalisierungsaufwand für eine Abmeldung ist genauso groß wie bei einer Anmeldung. Daher sind die Messungen der Gruppenanmeldungen ausreichend und auch für die Gruppenabmeldungen aussagekräftig.

- Bei den Gruppenadressen sind Sichtbarkeitsbegrenzungen (Scopes) im Modell vorgesehen. Das ermöglicht eine ausschließlich lokale Nutzung einer Gruppenadresse und erlaubt auch eine Mehrfachverwendung derselben Gruppenadresse in verschiedenen Netzsegmenten. Auf diese Begrenzung ist verzichtet worden, da sie keinen Einfluss auf das eigentliche Verfahren hat.

- Das reale ATM-Netz und die Netzsegmente mit den vorhandenen Controllern werden abstrahiert. Es wird nur noch ein Baum dargestellt, der die Hierarchie der

Controller angibt. Die Größe eines Netzes ist indirekt durch die wachsende Anzahl von Controllern repräsentiert und die Dichte eines Netzes, also die Anzahl der Endsysteme pro Controller ist äquivalent zu der Anzahl der Unterbäume.

Durch die hier genannten Vereinfachungen kann der Ansatz zur Gruppenverwaltung auf seine Kerneigenschaften reduziert werden. Um diese Eigenschaften auf ihre Skalierbarkeit untersuchen zu können, muss der Begriff Skalierbarkeit in Bezug auf die Verwaltung konkretisiert werden. Es werden drei Ausprägungen der Skalierbarkeit untersucht:

Gruppengröße: In welchem Maße wächst das Signalisierungsaufkommen und der Datenbestand in den Controllern mit steigender Gruppengröße.

Topologie: Welche Auswirkungen hat die geografische Ausdehnung eines Netzes auf den Verwaltungsaufwand einer Gruppe.

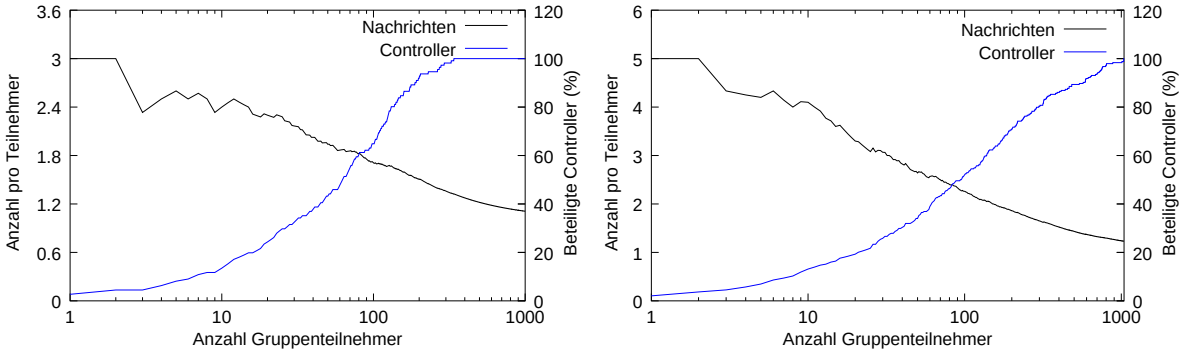
Anzahl der Gruppen: Sind in einem Netz viele Gruppen mit nur wenigen Teilnehmern angemeldet, kann dies auch zu einer hohen Belastung bei der Signalisierung führen.

Gemessen wird im Folgenden die mittlere Anzahl der Signalisierungsnachrichten pro Gruppenteilnehmer und die mittlere Anzahl der beteiligten Zwischensysteme (Controller). Des Weiteren werden die pro Controller bearbeiteten Nachrichten gemessen, um so ein Maß für die Belastung eines Controllers durch die Signalisierung zu erhalten.

Gruppengröße

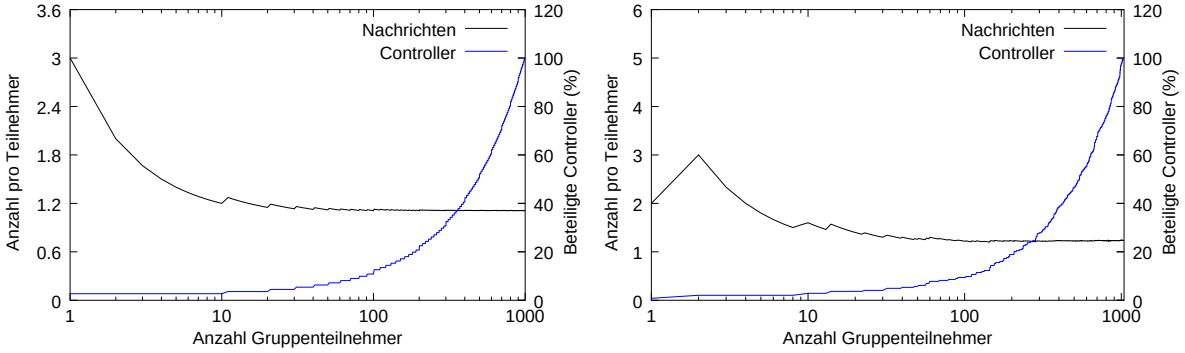
Die ersten durchgeführten Messungen beziehen sich auf die Gruppengröße. Hierzu ist das Verhalten einer einzelnen Gruppe untersucht worden. Einen vorläufigen Eindruck vermittelt Abbildung 6.31(a). Die Anzahl der Teilnehmer auf der x-Achse ist logarithmisch dargestellt, um eine bessere Auflösung bei kleinen Teilnehmergrößen zu erhalten. Es liegt der Messung ein vollständig ausgeglichener Baum mit 1000 Blättern (Endsystemen) zugrunde. Die Baumhöhe ist drei, und jeder Knoten im Baum hat exakt zehn Sohnknoten, insgesamt sind 111 Knoten (Controller) im Baum. Die Endsysteme melden sich in zufälliger Reihenfolge bei einer einzigen Gruppe an, bis alle 1000 Endsysteme der Gruppe angehören.

Die maximal mögliche Anzahl von Nachrichten pro Teilnehmeranmeldung ist durch die Höhe des Baums begrenzt. In Abbildung 6.31(a) ist zu erkennen, dass bei kleinen Gruppen dieses Maximum fast immer erreicht wird und bei steigender Gruppengröße die durchschnittliche Nachrichtenanzahl pro Teilnehmer gegen eins strebt. Die Kurve über die benötigten Nachrichten pro Teilnehmer ist durch die Aggregation der Teilnehmer in den Controllern begründet: Eine Nachricht wird nur weitergeleitet, wenn für die Gruppe noch keine Teilnehmer angemeldet sind. Je mehr Teilnehmer an einer Gruppe partizipieren, desto geringer werden die benötigten Nachrichten für einen neuen Teilnehmer. Als Minimum wird eine einzige Nachricht für die Anmeldung vom Endsystem beim lokalen Controller benötigt. Genau umgekehrt verhält es sich mit der Anzahl der an der Gruppe beteiligten Controller. Je mehr Gruppenteilnehmer angemeldet sind, desto mehr Controller sind auch in der Verwaltung dieser Gruppe involviert.



(a) ausgeglichener Baum, zufällige Anmeldung.

(b) zufälliger Baum, zufällige Anmeldung.



(c) ausgeglichener Baum, geordnete Anmeldung.

(d) zufälliger Baum, geordnete Anmeldung.

Abbildung 6.31.: Signalisierungsaufwand und Datenaufkommen bei steigender Gruppengröße, unterschiedlichen Baumtopologien und unterschiedlichen Teilnehmeranmeldeverhalten.

Eine andere Baumstruktur liegt Abbildung 6.31(b) zugrunde. Die Knoten des Baums wurden unter Berücksichtigung eines Zufallsfaktors konstruiert. Der Baum ist nicht ausgeglichen und die Baumhöhe schwankt bei den Endsystemen zwischen minimal zwei und maximal fünf Knoten. Durch die größere Baumhöhe ist die Anzahl der Nachrichten und Datensätze bei kleinen Gruppen höher als in Abbildung 6.31(a). Die Anzahl der Knoten (Controller) im Baum hat sich durch die unregelmäßige Struktur erhöht (im Vergleich zu Abbildung 6.31(a)) und beträgt hier 244 Knoten. Wie an den Kurven aber gut zu erkennen ist, hat die Baumstruktur nur geringe Auswirkungen auf das Verhalten des Ansatzes zur hierarchischen Gruppenverwaltung, der Kurvenverlauf ist annähernd identisch zu Abbildung 6.31(a).

Die Abbildungen 6.31(c) und 6.31(d) basieren auf denselben Bäumen wie Abbildung 6.31(a) bzw. 6.31(b). Geändert wurde hierbei die zufällige Auswahl der beitretenden Gruppenteilnehmer. Die Endsysteme treten den Gruppen in sequentieller Reihenfolge,

Teilbaum für Teilbaum, bei. Dieses, auf die Praxis bezogen, unnatürliche Vorgehen stellt den optimalen (minimalen) Fall bzgl. Nachrichtenanzahl und Datenbestand dar. Eine Besonderheit ist noch in Abbildung 6.31(d) zu erkennen. Durch die zufällige Knotenanordnung ist bei den ersten Teilnehmern die Entfernungen zur Wurzel und somit die Nachrichtenanzahl kleiner als die maximale Höhe des Baums.

Die maximale Anzahl von Nachrichten für eine Teilnehmeranmeldung ist direkt von der Baumhöhe abhängig und die Baumhöhe ist bei ausgeglichenen Bäumen logarithmisch von der Anzahl der Blätter abhängig. Bei N Endsystemen (Blättern) und jeweils m Sohnknoten pro Knoten errechnet sich die Baumhöhe h durch die Formel:

$$h = \lceil \log_m(N) \rceil$$

Bei nicht ausgeglichenen Bäumen kann die Entfernung zwischen Endsystem und Baumwurzel, die die maximale Nachrichtenanzahl pro Teilnehmeranmeldung bestimmt, zwischen minimal 1 und maximal N variieren.

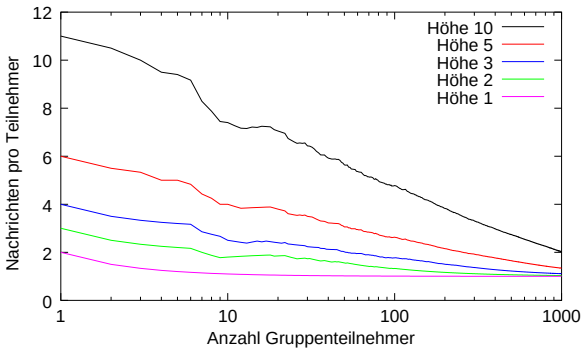
Die hier durchgeführten Messreihen zeigen eine gute Skalierbarkeit der Gruppenverwaltung bezüglich der Gruppengröße. Die Messreihen zeigen darüber hinaus deutlich, dass die Anzahl der Datensätze in den Controllern in direktem Zusammenhang mit der Nachrichtenanzahl steht. Die Anzahl der benötigten Nachrichten pro Teilnehmer ist dabei hauptsächlich von der Baumstruktur abhängig und nicht von der Gruppengröße. Damit ermöglicht dieses Schema eine effiziente Unterstützung großer Gruppen.

Topologie

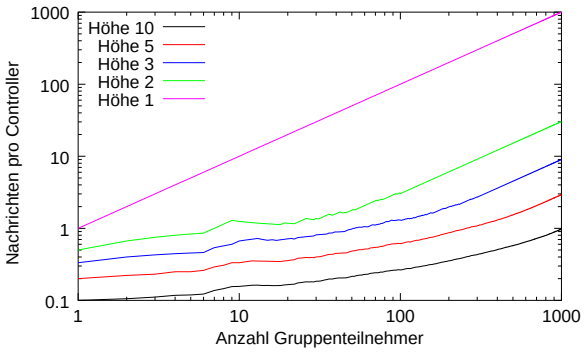
Die Topologie eines Netzes bzw. dessen geografische Ausdehnung hat insofern Einfluss auf die Skalierbarkeit, als die An- und Abmeldungen zu einer Gruppe verzögert werden können. Abgesehen von der Verzögerung, die durch die Struktur des Netzes vorgegeben und im Wesentlichen nicht zu beeinflussen ist, stellt die Anordnung der Controller einen wichtigen Faktor dar, der Einfluss auf die Anzahl der Nachrichten und die in den Controllern zu haltenden Datensätze hat. Das ist auch in den Abbildungen 6.31(a) und 6.31(b) zu erkennen. Bei dem zufällig organisierten Baum werden mehr Controller für die gleiche Anzahl Endsysteme benötigt als bei dem ausgeglichenen Baum und somit werden auch mehr Nachrichten versendet.

Bei ausschließlicher Betrachtung der Nachrichtenanzahl und der Datensätze in den Controllern wäre die optimale Konfiguration ein einziger zentraler Controller wie beim MARS (Unterkapitel 2.4, ab Seite 23). Hierbei wird aber die Auslastung des Controllers durch den Verwaltungs- und Signalisierungsaufwand außer acht gelassen. Die Anzahl der im Baum befindlichen Controller stellt also immer einen Kompromiss zwischen Nachrichtenanzahl und Belastung der Controller dar. Daher wird im Folgenden die Anzahl der gesendeten Nachrichten und die Anzahl der verarbeiteten Nachrichten pro Controller gegenübergestellt.

In Abbildung 6.32 sind diese beiden Kenngrößen gegenübergestellt. Die Messungen wurden für jeweils ca. 1000 Endsysteme (Blätter) mit unterschiedlichen Baumhöhen durchgeführt. Jeder Knoten im Baum hat dabei dieselbe Anzahl Sohnknoten. Den Messungen lagen die folgenden Baumstrukturen zugrunde:



(a) Versendete Nachrichten pro Teilnehmer.



(b) Verarbeitete Nachrichten pro Controller.

Abbildung 6.32.: Anzahl Nachrichten pro Teilnehmer vs. Nachrichten pro Controller bei unterschiedlichen Baumhöhen.

Baumhöhe (h)	Sohnknoten(g)	Blätter g^h	Innenknoten im Baum $(g^h - 1)/(g - 1)$
1	1000	1000	1
2	1024	32	33 (1 + 32)
3	1000	10	111 (1 + 10 + 100)
5	1024	4	341 (1 + 4 + 16 + 64 + 256)
10	1024	2	1023 (1 + 2 + 4 + ...+ 512)

Die Abbildung 6.32(a) zeigt, dass die mittlere Anzahl der pro Teilnehmer zu versendeten Nachrichten mit abnehmender Baumhöhe ebenfalls reduziert wird. Je weniger Controller durch die geringere Baumhöhe vorhanden sind, desto weniger Nachrichten werden auch generiert. Die Abbildung 6.32(b) gibt die Belastung der Controller wieder. Diese steht im umgekehrten Verhältnis zu Abbildung 6.32(a). Je mehr Controller beteiligt sind, desto weniger Nachrichten hat ein einzelner Controller zu verarbeiten. Zu beachten ist noch die logarithmische Darstellung der y-Achse in Abbildung 6.32(b). Im Diagramm ist die größte Differenz in der Belastung zwischen Baumhöhe 1 und Baumhöhe 2 zu erkennen. Daher sollte eine Organisation der Gruppenverwaltung in zwei Ebenen den größten Nutzen für eine Netzgröße von ca. 1000 Endsystemen darstellen.

Anzahl der Gruppen

Bei der Betrachtung der Gruppengröße und der Topologie ist immer nur eine einzige Gruppe aktiv gewesen. Es ist zu erwarten, dass bei mehr als einer Gruppe, die Anzahl der Daten und der Signalisierungsnachrichten linear mit der Anzahl der Gruppen ansteigt. Bei den bisher betrachteten Gruppen, deren Größe bis auf alle vorhandenen Endsysteme anstieg, sind somit alle Controller entsprechend der Anzahl der Gruppen mehr belastet. Anders sieht es aus, wenn es viele Gruppen mit relativ wenig Teilnehmern gibt.

Eine Gegenüberstellung der versendeten Nachrichten und der aktiven Controller, die

	Baum (Höhe 3)			Baum (Höhe 2)		
Anzahl Gruppen	Nachrichten	Controller		Nachrichten	Controller	
1	15	10,0	(9%)	10,7	5,7	(17%)
10	151,15	51,25	(46%)	106,65	26,45	(80%)
20	300,35	74,7	(67%)	214,8	31,35	(95%)
50	750,05	103,25	(93%)	535,05	33,0	(100%)
100	1502,25	110,4	(99%)	1071,0	33,0	(100%)
200	2998,45	111,0	(100%)	2139,8	33,0	(100%)

Tabelle 6.1.: Anzahl der versendeten Nachrichten und aktiven Controllern bei verschiedenen Gruppenanzahlen.

mindestens eine Gruppe verwalten, gibt Tabelle 6.1. Alle Gruppen bestehen immer aus fünf Teilnehmern. Die Gruppenteilnehmer sind dabei zufällig auf die Endsysteme verteilt und die Tabellenwerte repräsentieren die Mittelwerte aus 10 Simulationen. In der Tabelle 6.1 ist gut zu erkennen, wie die Anzahl der Nachrichten linear mit der Anzahl der Gruppen zunimmt. Damit nimmt auch die Anzahl der Datensätze zu, die in den Controllern gespeichert werden. Andererseits nimmt die Anzahl der beteiligten Controller nicht linear mit der Gruppenanzahl zu. Das ist dadurch zu erklären, dass die Anzahl der Controller zum einen begrenzt ist, wodurch ein Controller mehrere Gruppen verwaltet und zum anderen durch die zufällige Teilnehmerzuordnung zu den Endsystemen keine gleichmäßige Nutzung der Controller stattfindet.

Die Gruppenverwaltung zeigt eine gute Skalierbarkeit bezüglich der Gruppengröße und der Netztopologie. Die Anzahl der zu verarbeitenden Nachrichten hängt nicht von Anzahl der Teilnehmer, sondern von der Höhe des Verwaltungsbaumes ab. Der Signalisierungsverkehr steigt dagegen linear mit Anzahl der Gruppen an. Dies ist auch nicht anders zu erwarten, da die Gruppen separat voneinander behandelt werden müssen.

6.4.2. Bewertung des Datentransfers

In Unterkapitel 6.2 ist der Datentransfer zwischen den Teilnehmern einer Gruppe beschrieben worden. Für den Datentransfer wird, ähnlich wie bei der Gruppenverwaltung, eine Baumstruktur verwendet, die aber für jede Gruppe separat und je nach Bedarf dynamisch konstruiert wird.

Der entscheidende Faktor für die Effektivität des Datentransfers ist der Aufbau des Baumes und die Form desselben. Hieraus ergeben sich die Messreihen, die für den Datentransfer durchgeführt worden sind:

Anmeldeverzögerung: Die Zeitdifferenz zwischen einer Empfängeranmeldung und des ersten empfangenen Datenpaketes. Hierfür ist besonders die Höhe des Verwaltungsbaumes entscheidend.

MCS-Belastungen: Die Belastungen der MCS für den Baum einer Gruppe sind im

Wesentlichen alle identisch. Bei mehreren Gruppen und Bäumen ist es hingegen wünschenswert, dass unterschiedliche MCS genutzt werden, um Datenkonzentrationen in den einzelnen MCS zu vermeiden. Daher soll ermittelt werden, wie sich mehrere Gruppen im ATM-Netz anordnen.

Baum-Struktur: Der Baum für den Datentransfer einer Gruppe sollte im Idealfall möglichst wenig Zwischensysteme (MCS) beinhalten und kurze Datentransportwege zwischen den Teilnehmern ermöglichen. Es ist daher von Interesse zu beurteilen, inwieweit dieser Ansatz dieses Kriterium erfüllen kann.

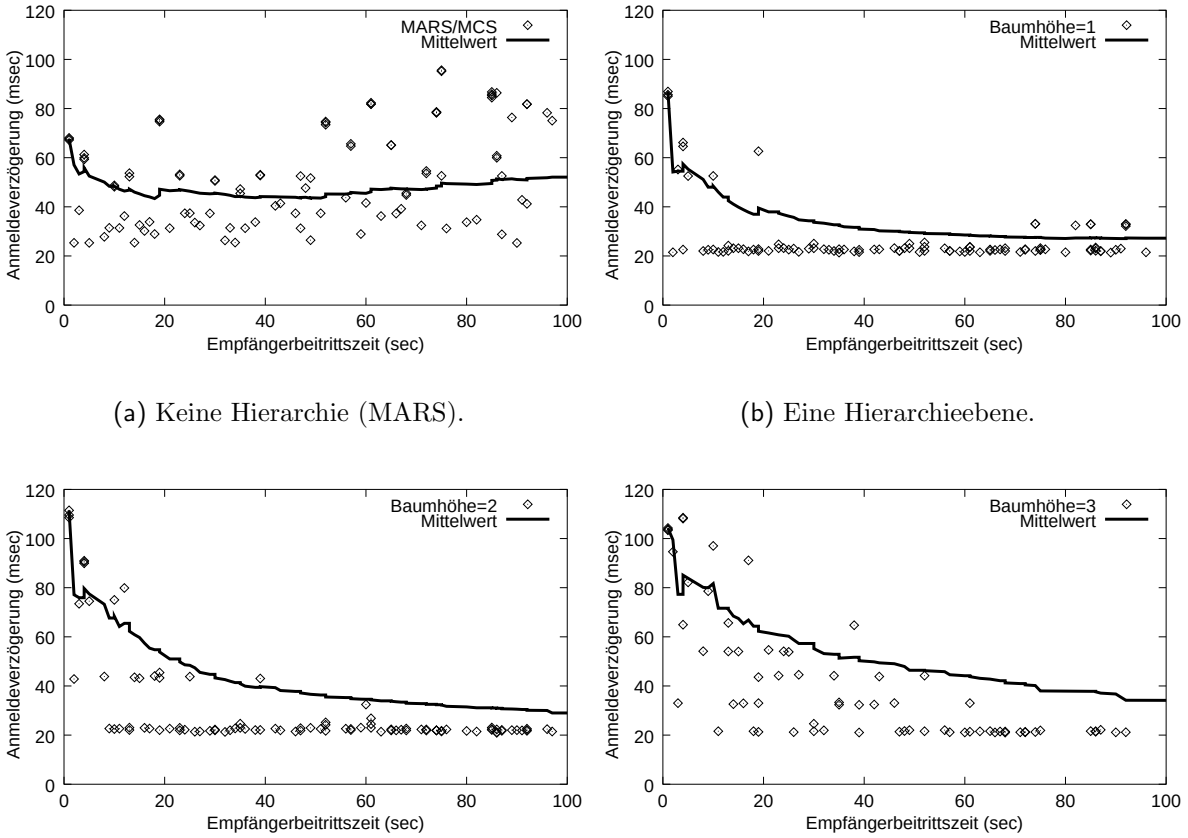
Anmeldeverzögerung

Bevor ein Empfänger nach seinem Gruppenbeitritt Daten empfangen kann, muss zuerst der Baum für den Datentransfer zu dem Empfänger aufgebaut bzw. erweitert werden. Hierzu ist eine Signalisierung erforderlich, die in Unterkapitel 6.2.2 ab Seite 100 erläutert worden ist. Nachdem mittels der Signalisierung alle notwendigen Informationen ausgetauscht worden sind, können die ATM-Verbindungen für den Anschluss des Empfängers aufgebaut werden.

Die Verzögerung, die bei einer Empfängeranmeldung entsteht, hängt im Wesentlichen davon ab, wie viele andere Empfänger sich schon in der Gruppe angemeldet haben und wieweit diese angemeldeten Empfänger vom neuen Empfänger entfernt sind. Die Entfernung hat hier eine doppelte Bedeutung, zum einen die geografische Entfernung, die Einfluss auf die Aufbaudauer der ATM-Verbindung hat und zum anderen die Entfernung im Verwaltungsbaum, denn jeder zusätzliche Controller bedeutet eine weitere Verzögerung bei der Signalisierung.

Für die Messreihe ist ein ATM-Netzwerk zugrunde gelegt worden, das in etwa der geografischen Verteilung des deutschen MBone-Netzes [61] entspricht. Jeder Controller benötigt eine konstante Zeit von 10 ms, um eine ankommende Signalisierungsnachricht zu verarbeiten, die Verzögerung der MCS entspricht den in Unterkapitel 5.3 ab Seite 67 spezifizierten Werten. Um die Anmeldeverzögerung messen zu können, ist zuerst ein zufällig ausgewählter Sender der Gruppe beigetreten und anschließend einhundert Empfänger in einem Zeitraum von einhundert Sekunden, bei denen die Verzögerungen gemessen worden sind. Für die Messreihen in Abbildung 6.33 sind Verwaltungsbäume mit einer unterschiedlichen Anzahl Ebenen verwendet worden. Die Abbildung 6.33(a) entspricht dem MARS/MCS-Schema (Unterkapitel 3.2) und bei den weiteren Abbildungen 6.33(b), 6.33(c) und 6.33(d) ist jeweils eine zusätzliche Hierarchieebene hinzugekommen. In Abbildung 6.33 sind die gemessenen Anmeldeverzögerungen eingezeichnet und der sich daraus errechnende Durchschnitt.

Die Abbildung 6.33(a) zeigt die Anmeldeverzögerung für das MARS/MCS-Schema. Die Verzögerung liegt im Mittel zwischen 40-60 ms für jeden Teilnehmer und setzt sich aus der Verzögerung im Controller und der Verzögerung beim ATM-Verbindungsaufbau zusammen. Bei dem SkaGAN-Ansatz in den Abbildungen 6.33(b) - 6.33(d) kann die Anmeldeverzögerung auf 20-30 ms reduziert werden. Wichtiger ist die genaue Quantifizierung der Zeitersparnis ist aber die prinzipielle Reduktion der Verzögerungszeiten, die



(a) Keine Hierarchie (MARS).

(b) Eine Hierarchieebene.

(c) Zwei Hierarchieebenen.

(d) Drei Hierarchieebenen.

Abbildung 6.33.: Verzögerung bei der Empfängeranmeldung bei unterschiedlichen Hierarchieebenen.

durch die Ausnutzung der Hierarchie erreicht werden kann. Während am Anfang noch hohe Verzögerungen entstehen, die auch deutlich höher als beim MARS oder MCS-Schema sind, reduziert sich die Verzögerung, je mehr Teilnehmer der Gruppe bereits beigetreten sind. Dieses Verhalten ist analog zu der Gruppenverwaltung, wie es z. B. in Abbildung 6.31(a) dargestellt ist. Je mehr Teilnehmer einer Gruppe beigetreten sind, desto größer ist die Wahrscheinlichkeit, dass bereits ein aktiver Teilnehmer in der Nähe des neuen Empfängers existiert und somit die ATM-Verbindung nur noch lokal etabliert werden muss.

Während bei einer Höhe des Verwaltungsbaumes von zwei bis drei für eine Anzahl von einhundert Teilnehmern in der Gruppe in den Abbildungen 6.33(b) und 6.33(c) die Anmeldeverzögerung sinkt, steigt in Abbildung 6.33(d) die Anmeldeverzögerung wieder an. Das ist durch einen erhöhten Verwaltungsaufwand zu erklären, der in keinem sinnvollen Verhältnis zur verwaltenden Gruppengröße steht. Jeder Controller hat hier weniger als fünf Endsysteme zu verwalten, was zu einem hohen Signalisierungsaufwand zwischen den Controllern in der Verwaltungshierarchie führt.

Insgesamt kann gesagt werden, dass die Anmeldeverzögerung durch die Einführung einer Hierarchie sowohl bei der Gruppenverwaltung als auch beim Datentransfer zu einer Reduktion der Anmeldeverzögerung führt. Die Vorteile machen sich für die Teilnehmer besonders bemerkbar, wenn bereits andere lokale Gruppenteilnehmer angemeldet sind, da dann die Anmeldeverzögerung signifikant reduziert werden kann. Die Vorteile der Hierarchie können aber auch durch zu viele Hierarchieebenen und dem hiermit verbundenen Verwaltungsaufwand im Verhältnis zur Gruppengröße negiert werden. Es sollte daher bei der Organisation des ATM-Netzes darauf geachtet werden, dass die entstehenden lokalen Einheiten nicht zu klein ausfallen, da ansonsten der Verwaltungsaufwand und die Anmeldeverzögerung steigt.

MCS-Belastungen

Gegenüber den Messungen in Unterkapitel 5.5, ab Seite 75, in dem die Belastung eines einzelnen MCS betrachtet worden ist, sollen in diesem Unterkapitel die Belastungen aller aktiven MCS im Netz analysiert werden. Insbesondere interessiert hierbei die gleichmäßige Lastverteilung auf verschiedene MCS. Die Grundlage ist diesmal kein lokales Netz, in dem der Controller die Lastverteilung auf mehrere MCS steuert, sondern ein Weitverkehrsnetz, in dem mittels des hierarchischen Gruppenkommunikationsschemas die Teilnehmer untereinander verbunden werden.

In Unterkapitel 6.2.1 ist das hierarchische Gruppenkommunikationsschema von SkaGAN erläutert worden und wie in diesem Schema die Lastverteilung möglich ist. Durch die Wahl unterschiedlicher primärer MCS für verschiedene Gruppen kann die Last der MCS in den höheren Ebenen gleichmäßig verteilt werden. Hierbei ist zu beachten, dass die Belastung aller aktiven MCS einer Gruppe gleich ist. Eine Ausnahme bilden nur die primären MCS für diese Gruppe, die durch die zusätzliche Verbindung zur nächst höheren Ebene eine etwas höhere Belastung haben. Eine Lastverteilung ist also nur zwischen mehreren Gruppen möglich. Eine Ausnahme bilden die in Unterkapitel 6.3 vorgestellten Erweiterungen, die aber erst im folgenden Unterkapitel 6.4.3 bewertet werden.

Für die Bewertung der Lastverteilung wird das in Abbildung 6.34 gezeigte Netzwerk zugrunde gelegt. Das Netzwerk ist in drei Ebenen gegliedert, und für jede Ebene ist immer ein Controller vorhanden. Die oberste Ebene ist ein Netz aus zwanzig Teilnetzen. Jedes dieser Teilnetze hat wiederum fünf weitere lokale Netze. Die lokalen Netze sind sternförmig aufgebaut und bestehen aus jeweils zehn Endsystemen und einem MCS. Insgesamt sind eintausend Endsysteme und einhundert MCS im Netzwerk vorhanden.

Die Abbildung 6.35 zeigt die Belastung der MCS für eine aktive Gruppe. Die Gruppe besteht aus insgesamt 20 Teilnehmern, 4 Sendern und 16 Empfängern, die zufällig auf die Endsysteme verteilt worden sind. Jeder Sender sendet mit einer Datenrate von 100 KBit/s. Von den hundert im Netz vorhandenen MCS werden 17 genutzt, wie die Abbildung 6.35 anhand der Spitzen gut erkennen lässt. Das bedeutet, dass einige Teilnehmer sich im selben lokalen Netz befinden und einen MCS gemeinsam benutzen. Die mittlere Belastung der MCS ist 0,0043 mit einer Varianz von $< \pm 6\%$.

Die nächste Abbildung 6.36 zeigt die Belastung der MCS für 20 aktive Gruppen im gleichen ATM-Netz. Die Gruppenzusammensetzung ist dabei identisch mit Abbildung

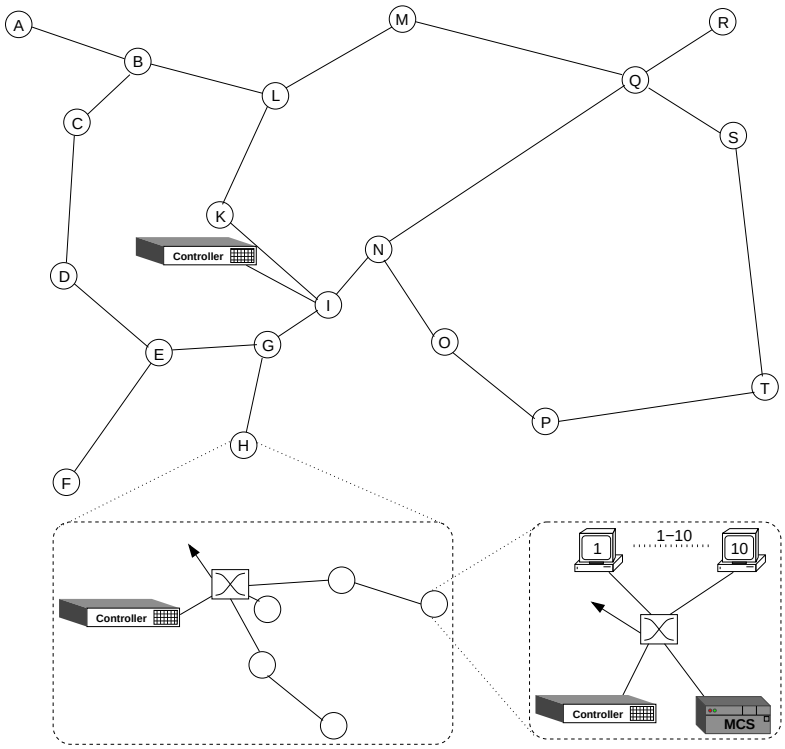


Abbildung 6.34.: Struktur des Netzwerkes für die Bewertung der Lastverteilung.

6.35. Um eine bessere Übersicht über die verschiedenen Belastungen der MCS zu bekommen, sind die gemessenen Werte aufsteigend sortiert. Die Zahlen in der Abbildung 6.36 geben an, in wie vielen Gruppen die MCS angemeldet sind. In der Summe sind bei allen MCS zusammen 354 Gruppen aktiv. Auf zwanzig Gruppen aufgeteilt, ergibt das im Mittel 17,7 aktive MCS pro Gruppe, was annähernd dem Wert von Abbildung 6.35 bei einer Gruppe entspricht. Im Mittel beträgt die MCS-Belastung 0,016, was etwa 3,8 aktiven Gruppen pro MCS bedeutet.

Wie in Abbildung 6.36 gut zu erkennen ist, sind die Differenzen der Belastungen zwischen den MCS bei mehreren Gruppen groß. Dennoch ist die höchste Verkehrskonzentration in einem MCS noch unter 50% des theoretisch möglichen Maximums. In einem MCS könnten in dem obigen Beispiel theoretisch maximal 20 Gruppen aktiv sein. Wie in Abbildung 6.36 zu erkennen ist, sind aber in der Simulation nur maximal 9 Gruppen in einem MCS aktiv. Das zeigt, dass eine gewisse Verteilung der Belastung auf die MCS erfolgt. Diese Verteilung ist allerdings nicht sehr effizient. Die Ursache hierfür liegt in der Wahl der primären MCS für eine Gruppe. Die MCS, deren Teilnehmer sich zuerst bei der Gruppe anmelden, werden als primäre MCS ausgewählt, ohne Berücksichtigung der Belastung. Da die Aufteilung der Teilnehmer auf die Endsysteme zufällig erfolgt, ist auch die Wahl der primären MCS vom Zufall abhängig.

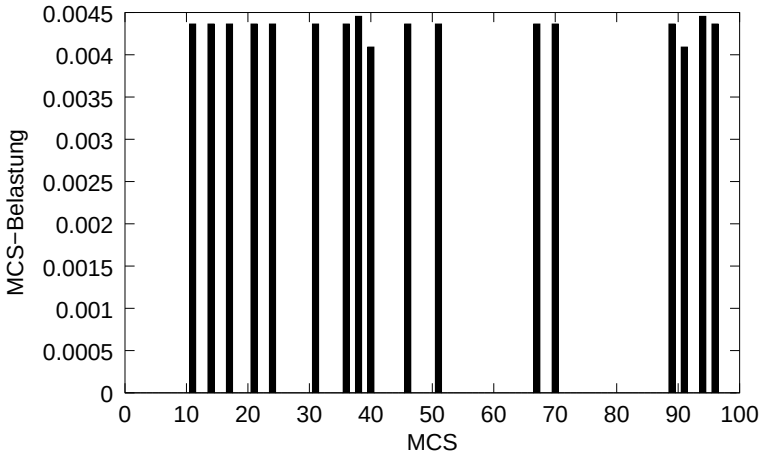


Abbildung 6.35.: MCS-Belastung für eine aktive Gruppe.

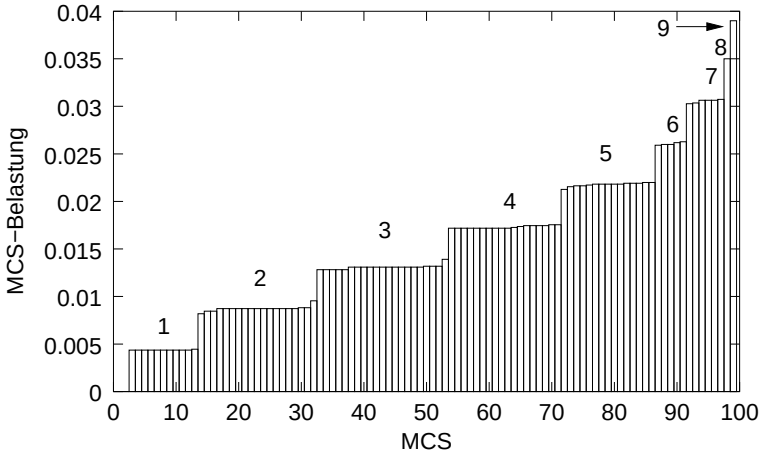


Abbildung 6.36.: MCS-Belastung für mehrere aktive Gruppen.

Baumstruktur

Bei der Messung der MCS-Belastungen sind nur die MCS betrachtet worden, ohne Berücksichtigung der Positionen im Netzwerk und in Bezug auf die Gruppen. Die Lage eines MCS und insbesondere die Lage der primären MCS hat aber einen entscheidenden Einfluss auf die Netzwerkauslastung. Befindet sich der höchste primäre MCS einer Gruppe z. B. am Rande des Netzwerks oder weit entfernt von den Gruppenteilnehmern, so führt das dazu, dass die Daten der Gruppenteilnehmer auf einer Leitung mehrfach übertragen werden müssen oder Leitungen unnötig verwendet werden.

Ein Beispiel hierzu zeigt Abbildung 6.37. Die Abbildung stellt ein kleines Netzwerk dar und drei Teilnehmer einer Gruppe, die alle gleichzeitig Sender und Empfänger sind. Alle Sender erzeugen das gleiche Datenvolumen, und die Pfeile geben an, wie viele Senderdaten auf der jeweiligen Leitung transportiert werden. Die Aufsummierung der Datenvolumen aller Leitungen ist dann ein Maß für das Datenaufkommen im Netzwerk für die Gruppe. Je geringer diese Summe ist, desto besser ist die Baumstruktur an die

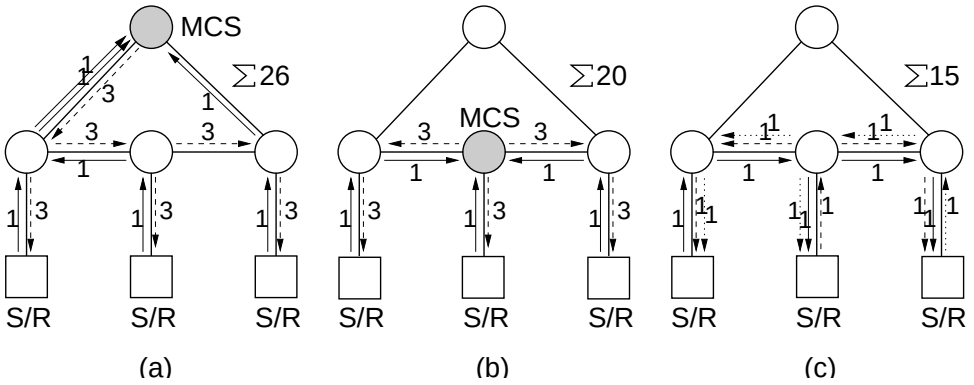


Abbildung 6.37.: Einfluss der Baumstruktur auf die Netzwerkbelastung: (a) nicht-optimale Position des MCS, (b) ideale Position des MCS, (c) VC Mesh (ohne MCS).

Netzwerktopologie und die Positionen der Teilnehmer angepasst.

Abbildung 6.37(a) zeigt das MARS/MCS-Schema, wobei die Lage des MCS ungünstig ist. Dieses Szenario erhält die schlechteste Bewertung der Belastung (Summe = 26). Hingegen zeigt Abbildung 6.37(b) die günstige Lage des MCS und die Belastung des Netzes reduziert sich um ca. 25% (Summe = 20). Eine minimale Belastung des Netzes zeigt Abbildung 6.37(c), der das VC-Mesh-Schema zugrunde liegt. Hier ist die Belastung um weitere 25% reduziert.

Optimierungskriterium:
Die Festlegung eines Minimums für die Netzbelastung erfordert ein Optimierungskriterium. Die Graphentheorie betrachtet Steiner-Bäume [62, 63] als diesbezüglich optimal. Steiner-Bäume sind Spannbäume mit minimalen Kosten. Aus zwei Gründen lassen sich Steiner-Bäume auf das hier vorliegende Problem nur sehr unzureichend anwenden:

1. Steiner-Bäume sind ungerichtet, daher werden symmetrische Verbindungen im Netzwerk vorausgesetzt und

2. unterschiedliche Senderdatenraten können nicht berücksichtigt werden, die Netzwerkressourcen werden global optimiert ohne Berücksichtigung der entstehenden Datenmengen auf den einzelnen Leitungen.

Des Weiteren kommt hinzu, dass die errechneten Steiner-Bäume nicht mit ATM realisiert werden könnten, da hierfür ATM Mehrpunkt-zu-Mehrpunkt-Verbindungen unterstützen müsste.

Eine andere Möglichkeit sind quellenbasierte Spannbäume, die z. B. mit dem Dijkstra-Algorithmus bestimmt werden können. Die ATM-Punkt-zu-Mehrpunkt-Verbindungen sind eine direkte Umsetzung dieser Spannbäume. Zwischen Quelle und Empfänger wird immer ein annähernd kürzester Weg etabliert. Es kann aber nicht garantiert werden, dass der Weg immer der kürzest mögliche ist, da hierzu eine globale Netzwerksicht notwendig wäre. Daher sind die ATM-Punkt-zu-Mehrpunkt-Verbindungen nur suboptimal.

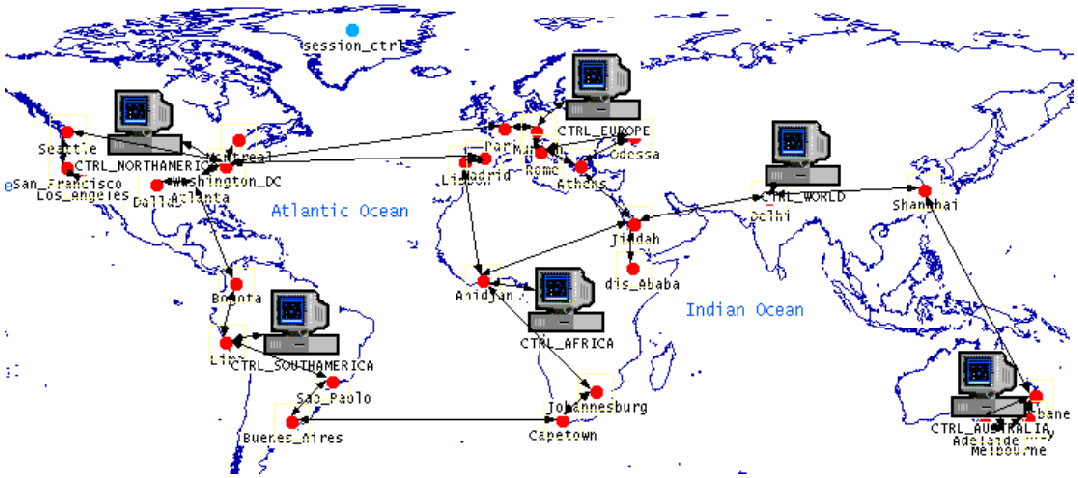


Abbildung 6.38.: Weitverkehrsszenario (fiktiv).

Eine Anwendung dieser Spannbäume stellt das VC-Mesh-Schema dar. Es wird mit dem VC-Mesh-Schema immer eine geringere Belastung erreicht als bei den MCS-basierten Schemata, da die hierfür zusätzlich benötigten Leitungen von/zum MCS für das VC-Mesh-Schema nicht erforderlich sind. Damit stellt das VC-Mesh-Schema immer eine untere Grenze dar, die angibt, wieweit die Netzwerkbelastung reduziert werden könnte.

Die auf diese Weise ermittelten Werte sind natürlich immer von der Netztopologie abhängig und können nicht auf andere Netztopologien übertragen werden. Dennoch ist hiermit eine Bewertung der Baumstruktur für den Datentransfer möglich.

Messungen:

Im Folgenden werden zwei Messreihen präsentiert. Der ersten Messreihe wird wieder das Netzwerk aus Abbildung 6.34 den Messungen zugrunde gelegt. Für die zweite Messreihe ist die in Abbildung 6.38 dargestellte Topologie gewählt worden. Jeder Knoten in Abbildung 6.38 repräsentiert dabei ein sternförmiges Netz mit jeweils 40 Endsystemem (siehe hierzu auch Anhang B.6, Seite 175). Die Messdaten werden für zwei Gruppengrößen erhoben. Es gibt eine kleine Gruppe, die identisch zur verwendeten Gruppe aus Abbildung 6.35 ist und aus 4 Sendern und 16 Empfängern besteht. Die große Gruppe besteht aus 4 Sender und 96 Empfängern. Gemessen wird, wie viele Leitungsabschnitte (virtual channel links, VCLs) pro Sender benötigt werden, damit die Daten an alle Empfänger ausgeliefert werden können. Werden dabei Daten auf einem VCL mehrfach übertragen, so wird dieser VCL auch entsprechend mehrfach gezählt. Es wird immer der Mittelwert und die Standardabweichung von jeweils 10 Messungen angegeben. Bei jeder Messung werden die zufälligen Positionen der Gruppenteilnehmer variiert.

Die Messungen werden für verschiedene Organisationsschemata durchgeführt. Als erstes werden die Schema des MARS-Ansatzes gemessen. Das ist zum einen das VC-Mesh-Schema gemessen, dessen Ergebnisse als unteres Minimum angenommen werden. Zum anderen das MCS-Schema, wobei für die erste Topologie (Abbildung 6.34) zwei unterschiedliche Positionen für den MCS gewählt worden sind. Der MCS befindet sich

	kleine Gruppe		große Gruppe	
Schema	VCLs	Standardabweichung	VCLs	Standardabweichung
VC Mesh	51,8	2,0 ($\pm 3,8\%$)	187,9	4,1 ($\pm 2,2\%$)
MCS zentral	57,3	2,8 ($\pm 4,9\%$)	193,3	3,9 ($\pm 2,0\%$)
MCS Rand	58,6	2,5 ($\pm 4,3\%$)	194,4	4,0 ($\pm 2,1\%$)
SkaGAN, 1 Ebene	71,7	4,1 ($\pm 5,2\%$)	222,9	12,1 ($\pm 5,4\%$)
SkaGAN, 2 Ebenen	78,2	4,6 ($\pm 5,6\%$)	254,0	14,9 ($\pm 5,8\%$)

Tabelle 6.2.: Bewertung verschiedener Baumstrukturen für zwei unterschiedlichen Gruppengrößen

	kleine Gruppe		große Gruppe	
Schema	VCLs	Standardabweichung	VCLs	Standardabweichung
VC Mesh	36,3	1,8 ($\pm 4,8\%$)	124,3	2,0 ($\pm 1,6\%$)
MCS	41,2	2,5 ($\pm 5,9\%$)	126,6	2,3 ($\pm 1,8\%$)
SkaGAN, 2 Ebenen	53,6	3,1 ($\pm 5,7\%$)	145,8	8,1 ($\pm 5,5\%$)

Tabelle 6.3.: Bewertung verschiedener Baumstrukturen bei einem Weitverkehrsszenario.

einmal in der Netzmitte (Knoten 'I' in Abbildung 6.34) und einmal am Netzrand (Knoten 'R' in Abbildung 6.34). Bei der zweiten Topologie (Abbildung 6.38) befindet sich der MCS bei dem Knoten 'Athen'. Die verbleibenden Messungen werden mit dem hierarchischen Schema von SkaGAN durchgeführt, einmal mit einer, und einmal mit zwei Ebenen.

Messergebnisse:
Die Ergebnisse der Messungen sind in den Tabellen 6.2 und 6.3 dargestellt. Bei allen Werten ist die Standardabweichung unter 6%, die Unterschiede bei der Verteilung der Teilnehmer auf die Endsysteme hat also nur einen geringfügigen Einfluss auf die Ergebnisse. Die erhaltenen Werte hängen sehr stark mit der Netzwerktopologie zusammen. Beim VC-Mesh-Schema bedeutet das Ergebnis z. B., dass in der kleinen Gruppe pro Sender die Daten auf 51,8 Leitungen versendet werden, damit alle Empfänger die Daten erhalten können. In der großen Gruppe werden ca. dreimal so viele Leitungen benötigt. Das ist insofern günstig, da die Empfängeranzahl hier um das sechsfache gestiegen ist.

Beim MCS-Schema liegen die Ergebnisse etwa 10-15% höher als beim VC-Mesh-Schema. Der Leitungsbedarf zwischen kleiner und großer Gruppe ist auch etwas stärker angestiegen als beim VC-Mesh-Schema. Das ist durch den notwendigen MCS zu erklären, der eine Konzentration der Daten erzwingt. Hierdurch nehmen die Datenpakete nicht mehr die kürzesten Wege zwischen Sender und Empfänger, im Gegensatz zum VC-Mesh-Schema. Interessanter ist jedoch die Tatsache, dass sich der Leitungsbedarf bei beiden MCS-Schema kaum unterscheidet und beide Werte innerhalb des jeweiligen Toleranzbereiches liegen. Das bedeutet, dass der Leitungsbedarf unabhängig von der Po-

sition des MCS im Netz ist. Diese Aussage ist in diesem Falle zutreffend, ist aber nicht für jeden Fall richtig. Zwei Ursachen haben das fast identische Verkehrsaufkommen zur Folge:

1. Nicht für alle Sender wird die Entfernung zum MCS länger, wenn dieser sich an einer ungünstigen Position im Netz befindet, die Entfernung kann sich auch reduzieren.

2. In der Gruppe sind wesentlich mehr Empfänger als Sender, und die Verteilung der Daten vom MCS an die Empfänger berücksichtigt wieder die kürzesten Wege auf einer Punkt-zu-Mehrpunkt-Verbindung. Damit werden die Datenpakete im ATM-Netz an günstig gelegenen Knotenpunkten dupliziert.

Beide Ursachen zusammen bewirken, dass das Datenaufkommen im Netz auch bei einer ungünstigen MCS Position nur marginal ansteigt.

Der bei SkaGAN verwendete Ansatz mit einer hierarchischen Baumstruktur hat noch einen höheren Bandbreitenbedarf, wie aus Tabelle 6.2 und 6.3 gut zu erkennen ist. Gegenüber dem VC-Mesh-Schema ist der Bedarf bei der kleinen Gruppe um 51% und bei der großen Gruppe um 35% angestiegen. Das ist zum einen durch den zusätzlichen Datenaustausch zwischen den MCS bedingt und zum anderen durch die nicht immer optimalen Kommunikationswege. Da die Wahl der primären MCS nur durch die zeitliche Anmeldereihenfolge bestimmt wird, kann es hierbei zu ungünstigen Konstellationen kommen, die dazu führen, dass auf einigen Leitungen Daten mehrfach übertragen werden. Hier ist auch der Unterschied zwischen dem Baum mit einer Ebene und dem Baum mit zwei Ebenen begründet. Die Anordnung in der ersten Messreihe der MCS und der Controller für den Baum mit zwei Ebenen zeigt Abbildung 6.34. Für den Baum mit einer Ebene sind die Controller und MCS nur jeweils in der mittleren Ebene vorhanden und nicht mehr in den lokalen Netzen. Hierdurch werden im Vorhinein schon einmal eine Reihe von ungünstigen MCS Positionen im Netz vermieden, was zu einem geringeren Bandbreitenbedarf führt.

Dem erhöhten Leitungsbedarf des bei SkaGAN benutzten Baumschemas steht im Vergleich zum MCS-Schema eine geringere Verzögerung gegenüber. Das ist in Abbildung 6.39 dargestellt. Die Verzögerungen wurden im Szenario aus Abbildung 6.34 gemessen. In Abbildung 6.39 sind die verschiedenen Schemata und Gruppengrößen dargestellt, sortiert nach dem aufsteigenden Mittelwert. Zusätzlich sind die Standardabweichung und die gemessenen Minima und Maxima eingezeichnet. Trotz der erhöhten Anzahl an MCS ist die durchschnittliche Ende-zu-Ende-Verzögerung bei SkaGAN niedriger als beim MCS-Schema. Das ist darin begründet, dass lokale bzw. nah gelegene Gruppenteilnehmer die Daten der Sender auf einem entsprechend kürzeren Wege erhalten. Es müssen nicht alle Datenpakete den zentralen MCS passieren, wie beim MCS-Schema.

6.4.3. Bewertung der Erweiterungen

In Unterkapitel 6.3, ab Seite 113 sind zwei Verfahren vorgeschlagen worden, wie eine verbesserte Lastverteilung bei SkaGAN erreicht werden kann. Das erste Verfahren in

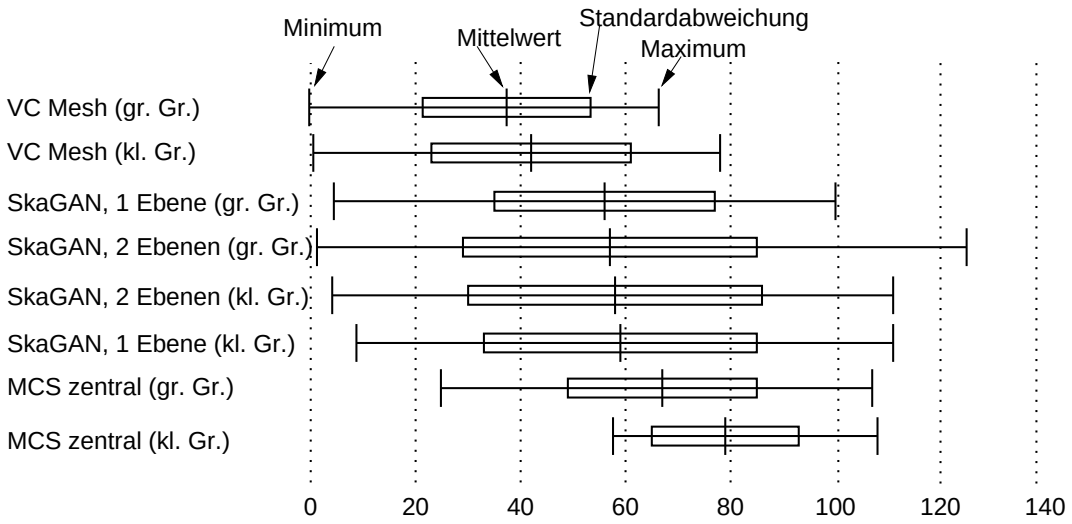


Abbildung 6.39.: Ende-zu-Ende-Verzögerung (in Millisekunden) der verschiedenen Schemata.

Unterkapitel 6.3.1, ab Seite 114 behandelt die Ersetzung eines primären MCS. Damit ist es möglich, MCS an ungünstigen geografischen Positionen auszutauschen, und so die Netzwerkbelastung zu reduzieren. Das dritte Verfahren (Unterkapitel 6.3.2, ab Seite 119) behandelt die Nutzung mehrerer paralleler Bäume für eine Gruppe. Damit können besonders große Gruppen oder Gruppen mit einem hohen Datenvolumen besser unterstützt werden.

Ersetzung eines primären MCS

Für die Ersetzung eines primären MCS durch einen anderen MCS sind vier Messreihen durchgeführt worden:

- Messung der Datenraten bei den MCS und den Empfängern. An einem kleinen Beispiel soll gezeigt werden, dass das Verfahren funktioniert und dass es die gewünschten Resultate erzeugt.
- Messung der Ende-zu-Ende-Verzögerung. Durch die MCS-Ersetzung kann in der Regel die Anzahl der beteiligten MCS reduziert werden. Daher ist durch dieses Verfahren eine Reduzierung der Ende-zu-Ende-Verzögerung zu erwarten.
- Unterbrechungsdauer der Datenübertragung während der MCS-Ersetzung.
- Die Häufigkeit der MCS-Ersetzungen ist ebenfalls von Interesse. Bei jeder Ersetzung ist die Kommunikation der betroffenen Gruppe unterbrochen, daher sollten möglichst wenig Ersetzungen stattfinden.

Für die Messreihen bei der MCS-Ersetzung ist als auslösendes Ereignis nur der Austritt aller lokalen Teilnehmer modelliert. Wenn kein lokaler Teilnehmer mehr in der Gruppe

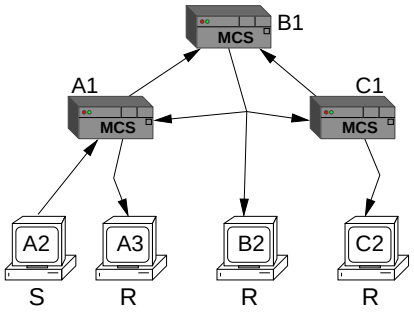


Abbildung 6.40.: Gruppenkommunikationsschema mit drei lokalen Netzen.

existiert, verlässt der MCS die Gruppe ebenfalls. Die MCS-Ersetzung aufgrund einer erhöhten MCS-Belastung ist hier nicht weiter untersucht worden. Die Ergebnisse der Messungen können aber auch auf diesen Fall angewendet werden.

Messung der Empfängerdatenraten:
Zunächst wird der Nachweis erbracht, dass das in Unterkapitel 6.3.1 beschriebene Verfahren zur Ersetzung eines primären MCS auch funktioniert. Hierzu wird die Anzahl der empfangenen Bits in den Empfängern gezählt, einmal mit und einmal ohne MCS-Ersetzung. Das hierfür verwendete Gruppenkommunikationsschema zeigt Abbildung 6.40.

Dargestellt ist ein Gruppenkommunikationsschema in zwei Ebenen mit drei lokalen Netzen A, B und C. In jedem lokalen Netz ist ein MCS und ein empfangender Teilnehmer vorhanden. Außerdem existiert ein Sender im lokalen Netz A. Der MCS B1 ist primärer MCS der Ebene 2, die beiden anderen MCS sind primäre MCS der Ebene 1.

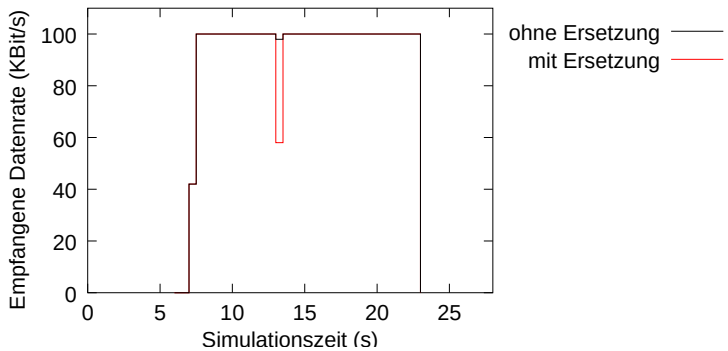


Abbildung 6.41.: Empfangene Datenrate des Endsystems C2.

Die Abbildung 6.41 zeigt exemplarisch die empfangene Datenrate für Empfänger C2 mit und ohne MCS-Ersetzung. Empfänger C2 tritt bei Zeitpunkt 7 s der Gruppe bei und verlässt sie bei 24 s. Der Sender A2 ist in dieser Zeit immer aktiv. Er sendet konstant mit 100 KBit/s (100 Pakete pro Sekunde mit einer Größe von 1000 Bits).

Bei 13 s tritt der Teilnehmer B2 im lokalen Netz aus. Bei aktivierter MCS-Ersetzung wird der MCS B1 durch den MCS C1 ersetzt (MCS A1 wäre auch möglich), ansonsten

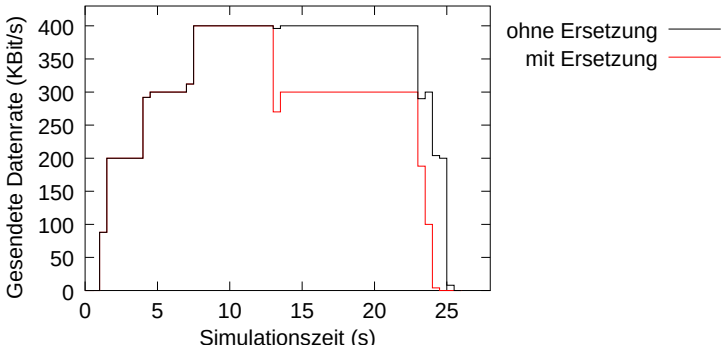


Abbildung 6.42.: Gesendete Datenrate der MCS mit und ohne MCS-Ersetzung.

bleibt der MCS B1 aktiv. Dadurch entsteht eine Unterbrechung in der Gruppenkommunikation, wie Abbildung 6.41 zeigt. Diese Unterbrechung ist bedingt durch die Implementierung der Punkt-zu-Mehrpunkt-Verbindungen im OpNet Simulator. Während des Zeitraums, in dem einer ATM-Verbindung ein Teilnehmer hinzugefügt oder entfernt wird, können auf dieser ATM-Verbindung keine Daten gesendet werden. Im dargestellten Graphen ist keine komplette Unterbrechung zu erkennen, da hier die empfangene Datenrate pro Zeitintervall summiert und nur der Durchschnitt angegeben wird. Im Fall ohne Ersetzung wird die Verbindung von MCS B1 nach Endsystem B2 abgebaut, was zu einem kleinen Einbruch in der Datenrate führt. Bei der MCS-Ersetzung entsteht eine längere Unterbrechung, weil mehrere Verbindungen geändert werden müssen. Das ist deutlich am Verlauf der gestrichelten Linie in Abbildung 6.41 zu erkennen. Ansonsten ist zu sagen, dass die empfangenen Daten sowohl ohne als auch mit MCS-Ersetzung identisch sind (außer bei der Unterbrechung), da beide Graphen direkt übereinander liegen.

Die beiden Kurven in Abbildung 6.42 zeigen die Summe der gesendeten Datenraten aller drei MCS. Der stufenweise Anstieg der gesendeten Datenrate bei den MCS entsteht jeweils durch die zeitlich verzögerten Beitritte der Teilnehmer. Genau umgekehrt verhält es sich mit den Austritten der Teilnehmer am Ende der Simulationszeit. Der Austritt des Empfängers B2 bei 13 s löst die MCS-Ersetzung aus, falls diese aktiviert ist. An der Kurve in Abbildung 6.42 ist zu erkennen, dass die Datenrate bei der MCS-Ersetzung geringer ist. Das ist dadurch zu erklären, dass die Daten nicht mehr über den überflüssigen MCS B1 gesendet werden müssen und somit ein Zwischensystem entfällt.

Messung der Ende-zu-Ende-Verzögerung:
Im Folgenden soll die durch die MCS-Ersetzung verursachte Änderung der Ende-zu-Ende-Verzögerung untersucht werden. Das hierzu verwendete Szenario besteht aus nur vier lokalen Netzen. In jedem lokalen Netz ist ein Endsystem und ein MCS vorhanden. Damit die Verzögerungen hervorgehoben werden können, wird ein geografischer Kontext verwendet. Die lokalen Netze sind mehrere 1000 Kilometer voneinander entfernt und auf verschiedene Kontinente verteilt worden, wie Abbildung 6.43 zeigt.

In der Gruppe sind alle Endsysteme als Empfänger angemeldet und ein Sender ist aktiv. Dieser Sender befindet sich im lokalen Netz C in Australien. Im lokalen Netz A in

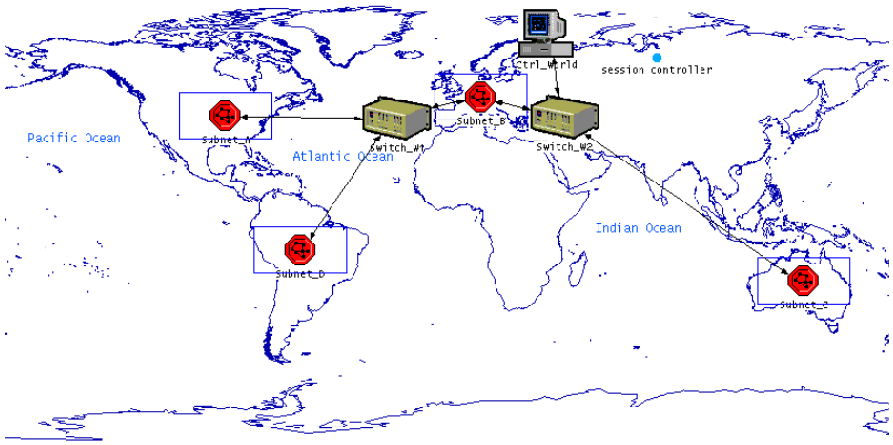


Abbildung 6.43.: ATM-Netz verteilt über vier Kontinente.

Endsystem	Verzögerung (ms)	
	vorher	nachher
A	80	—
B	108	51
C	6,6	6,6
D	122	85

Tabelle 6.4.: Ende-zu-Ende-Verzögerung vor und nach der MCS-Ersetzung.

den USA befindet sich der primäre MCS. Dieser lokale Empfänger tritt aus der Gruppe aus, wodurch eine MCS-Ersetzung ausgelöst wird. Der Ersatz-MCS ist in diesem Fall der MCS im lokalen Netz B in Europa. Die Tabelle 6.4 zeigt die Verzögerungen für das Beispiel vor und nach der MCS-Ersetzung. Beim Sender C, der auch Empfänger ist, ändert sich die Verzögerung nicht. Bei den Endsystemen B und D verringert sich die Verzögerung, da die Datenpakete jetzt eine insgesamt kürzere Entfernung und ein Zwischensystem weniger zu passieren haben.

Es ist aber ebenso gut möglich, dass die Verzögerung durch eine MCS-Ersetzung erhöht wird. Wenn ein MCS eine Gruppe verlässt, der sich an einer zentralen (günstigen) Position im Netz befand, so muss zu einem anderen, ungünstiger gelegenen MCS gewechselt werden und die Verzögerung erhöht sich.

Unterbrechungsdauer während der MCS-Ersetzung:
Ein anderer wichtiger Faktor, der hier kurz aufgezeigt werden soll, ist die Dauer der Unterbrechung während der Ersetzung eines primären MCS. Diese Unterbrechung ist nicht zwingend, aber in der aktuellen Version des Simulationsmodells vorhanden, da zuerst immer die bestehenden Verbindungen abgebaut werden, bevor die neuen Verbindungen aufgebaut werden (siehe hierzu auch Unterkapitel 6.3.1, ab Seite 114). Die gemessene Unterbrechungsdauer beträgt im Mittel ∼200 ms und setzt sich aus drei Zeiträumen zusammen: Signalisierung zwischen Controllern und MCS, ATM-Signalisierung und die

Verzögerung bei der Versendung der Dateneinheiten. Die Tabelle 6.5 zeigt einen typischen Ersetzungsvorgang und wie sich die Zeiträume aus den einzelnen Schritten bei der MCS-Ersetzung ergeben. Die erste Spalte der Tabelle 6.5 gibt dabei die relativen Zeitpunkt der einzelnen Aktionen seit Beginn der Ersetzung an. Die zweite Spalte gibt die Dauer der Unterbrechung des Datentransportes während der MCS-Ersetzung an.

Zeitpunkt (ms) Signalisierung	Unterbrechung (ms) Datentransport	Beschreibung	Signalisierung
0	–	Ersatz-MCS ist ermittelt	Controller
45,1	–	Aktualisierungen der Controller	
48,6	0	Verbindungsabbau bei MCS B1	
50,2	1,6	Verbindungsabbau bei MCS C1	
50,4	1,8	Verbindungsaufbau bei MCS C1	
50,9	2,3	Verbindungsabbau bei MCS A1	
51,1	2,5	Verbindungsaufbau bei MCS A1	
162,1	113,5	Verbindung bei MCS A1 akzeptiert	ATM
162,9	114,3	Verbindung bei MCS C1 akzeptiert	
209,4	160,8	Verbindung bei MCS C1 aufgebaut	
210,1	161,5	Verbindung bei MCS A2 aufgebaut	
255,1	206,5	Verzögerung der Dateneinheiten	keine

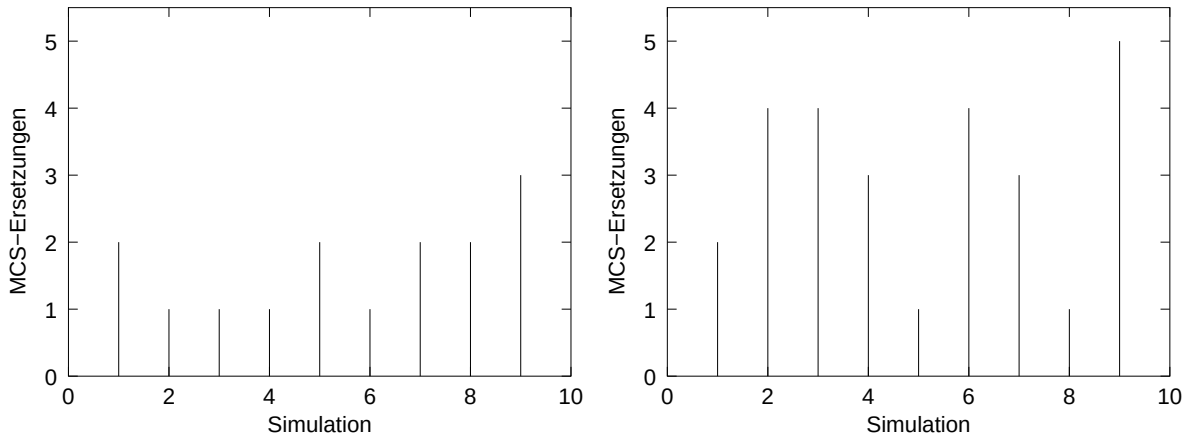
Tabelle 6.5.: Zusammensetzung der Unterbrechungsdauer bei einer MCS-Ersetzung.

Die Gesamtdauer und die Unterbrechungsdauer sind verschieden. Während der ersten Phase, in der ein Ersatz-MCS ermittelt wird, ist der zu ersetzenden MCS noch aktiv. Daher setzt sich die Unterbrechungsdauer fast ausschließlich aus dem Zeitraum für den ATM-Verbindungsabbau und -aufbau zusammen.

Häufigkeit der MCS-Ersetzungen:
Um abschätzen zu können, wie oft Unterbrechungen auftreten, muss die Anzahl der MCS-Ersetzungen abgeschätzt werden. Der ungünstigste Fall entsteht hierbei, wenn die Gruppenzusammensetzung so gegeben ist, dass in jedem lokalen Netz genau ein Teilnehmer vorhanden ist.

Wird dabei (zufällig) immer der Ersatz-MCS so gewählt, dass dieser in dem lokalen Netz des Teilnehmers ist, der als Nächster die Gruppe verlässt, so treten bei N Teilnehmern auch N MCS-Ersetzungen auf. Jedoch ist diese Kombination sehr unwahrscheinlich.

Allgemein ist zu beobachten, dass zwischen der Dynamik einer Gruppe und der Anzahl der MCS-Ersetzungen ein enger Zusammenhang besteht. Dieses hängt aber von der Reihenfolge der Austritte der Teilnehmer und der Wahl des Ersatz-MCS ab. Daher sollen im Folgenden zufällige Gruppenzusammensetzungen betrachtet werden. Alle Gruppen bestehen jeweils aus acht Teilnehmern, wobei jeweils die eine Hälfte Sender und die andere Hälfte Empfänger sind. Das zugehörige Netz besteht aus acht lokalen Teilnetzen mit jeweils einem Endsystem und ist in zwei Ebenen gegliedert. Diese Annahme ist nicht sehr praxisnah, aber für die Ersetzung eines MCS interessiert nur, ob er kein oder



(a) Baum mit 2 Ebenen.

(b) Baum mit 3 Ebenen.

Abbildung 6.44.: Anzahl der MCS-Ersetzungen.

noch mindestens ein lokales Endsystem versorgt. Daher wird mit diesem Szenario bei der Abmeldung eines Endsystems aus einer Gruppe immer auch eine MCS-Ersetzung eingeleitet.

Die Abbildung 6.44 zeigt die Anzahl der Restrukturierungen. Insgesamt sind zehn Simulationen mit zufälligen Bei- und Austrittszeiten der Teilnehmer durchgeführt worden. Der Mittelwert beträgt beim Szenario mit zwei Baumebenen 1,7 MCS-Ersetzungen (Abbildung 6.44(a)). Die Abbildung 6.44(b) zeigt die Anzahl der MCS-Ersetzungen in einem Baum mit 3 Ebenen bei den gleichen Gruppenzusammensetzungen und -änderungen. In diesem Baum werden jeweils drei bzw. einmal zwei lokale Netze zu einem Netzsegment der Ebene 2 zusammengefasst. Auf Ebene 3 werden die drei Netzsegmente der Ebene 2 zusammengefasst. Abbildung 6.44 zeigt deutlich, dass bei dem Austritt eines MCS häufig mehr als eine MCS-Ersetzung durchgeführt werden müssen.

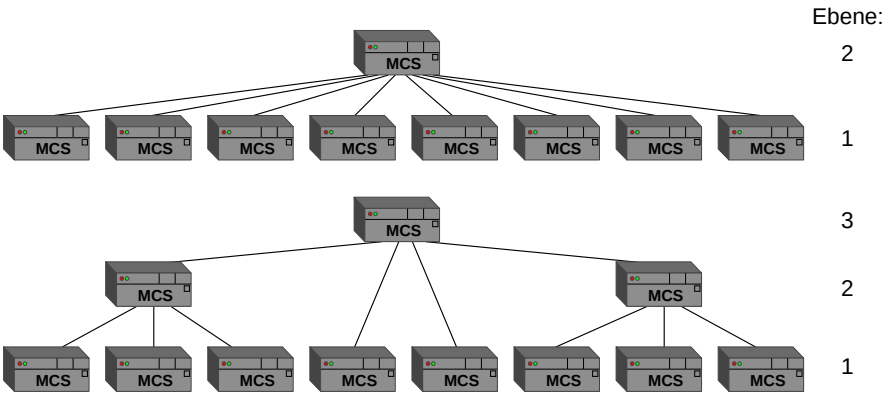


Abbildung 6.45.: Primäre MCS bei unterschiedlichen Hierarchieebenen.

Die Anzahl der MCS-Ersetzungen bei 3 Ebenen ist im Vergleich zum Szenario mit

zwei Ebenen deutlich höher. Der Mittelwert beträgt 3,1 MCS-Ersetzungen. D. h. die Anzahl der MCS-Ersetzungen steigt mit der Anzahl der Hierarchieebenen. Dieses ist dadurch zu erklären, dass mehr primäre MCS in Betracht kommen, für die eine MCS-Ersetzung durchgeführt werden kann bzw. muss. Die Abbildung 6.45 verdeutlicht dieses Verhalten.

In Abbildung 6.45 sind nur die MCS auf den einzelnen Hierarchieebenen dargestellt. Beim oberen Szenario wird eine Hierarchie mit zwei Ebenen verwendet, in der nur ein primärer MCS der Ebene 2 vorhanden ist. Im unteren Szenario sind die gleichen MCS vorhanden, nur sind sie in einer Hierarchie mit drei Ebenen angeordnet, mit zwei primären MCS der Ebene 2 und einem primären MCS der Ebene 3. Die Abbildung verdeutlicht, dass bei mehr Ebenen die MCS-Ersetzungen häufiger auftreten können. Zum einen sind bei mehr Ebenen i. Allg. mehr MCS pro Endsystem vorhanden und zum anderen müssen häufiger mehrere rekursive Ersetzungen durchgeführt werden, um einen primären MCS aus dem Baum zu entfernen.

Bei einem Baum mit mehreren Ebenen sind in der Regel häufiger und mehr MCS-Ersetzungen zu erwarten. Demgegenüber steht aber der Vorteil, dass viele MCS-Ersetzungen nur lokal in einer unteren Ebene ausgeführt werden. Da jede MCS-Ersetzung auch eine Unterbrechung beim Datentransfer verursacht, sind bei Bäumen mit mehreren Ebenen die Unterbrechungen auch häufiger lokal und somit ist immer nur eine verhältnismäßig kleine Teilgruppe von der Unterbrechung betroffen.

Bei einer Struktur mit mehreren Ebenen kommen auch insgesamt mehr MCS-Ersetzungen vor. Diese Ersetzungen beschränken sich dafür aber meist auf ein begrenzten Teil der Gruppe, wodurch die Unterbrechung der Gruppenkommunikation ebenfalls begrenzt werden kann.

Lastverteilung durch parallele Bäume

Um die Nutzung mehrerer Bäume für eine Gruppe bewerten zu können, sind im Folgenden drei verschiedene Messreihen durchgeführt worden:

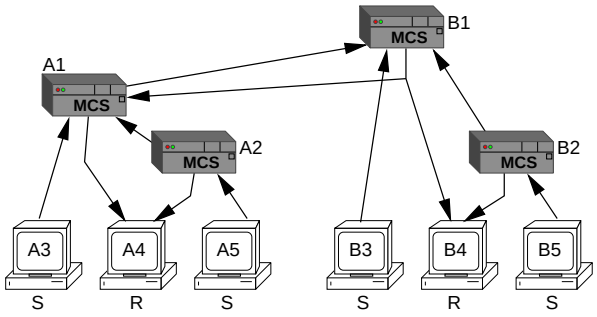
- Messung der Datenraten bei den Empfängern und den MCS. Hiermit wird der Nachweis der Funktionsfähigkeit des in Unterkapitel 6.3.2 beschriebenen Verfahrens erbracht.

- Dauer der Datentransportunterbrechung durch den Aufbau eines neuen Baumes. Wie bei der MCS-Ersetzung entsteht auch beim Aufbau eines neuen Baumes für einen Teil der Empfänger eine Unterbrechung im Datenfluss.

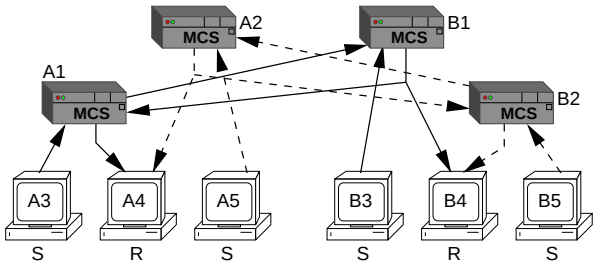
- Belastungsverteilung auf die MCS. Das Ziel von parallelen Bäumen innerhalb einer Gruppe ist, die Belastung gleichmäßiger auf die MCS verteilen zu können und somit Datenkonzentrationen zu vermeiden.

Bei der Untersuchung der Lastverteilung durch parallele Bäume wird im Folgenden immer nur der Aufbau eines neuen Baumes betrachtet. Der Abbau eines Baumes wird nicht weiter beachtet.

Messung der Empfängerdatenraten:
Wie bei der MCS-Ersetzung wird beim Aufbau eines bzw. mehrerer Bäume zuerst geprüft, ob die empfangenen Datenraten bei den Empfängern mit und ohne Aufbau eines neuen Baums identisch sind. Das verwendete Szenario besteht aus zwei lokalen Netzen A und B, in denen jeweils zwei MCS und drei Endsysteme vorhanden sind, wovon ein Endsystem Empfänger und zwei Endsysteme Sender sind. Die Hierarchie besteht aus zwei Ebenen. Die beiden Szenarien mit einem und mit zwei Bäumen sind in Abbildung 6.46 dargestellt.



(a) Ein Baum.

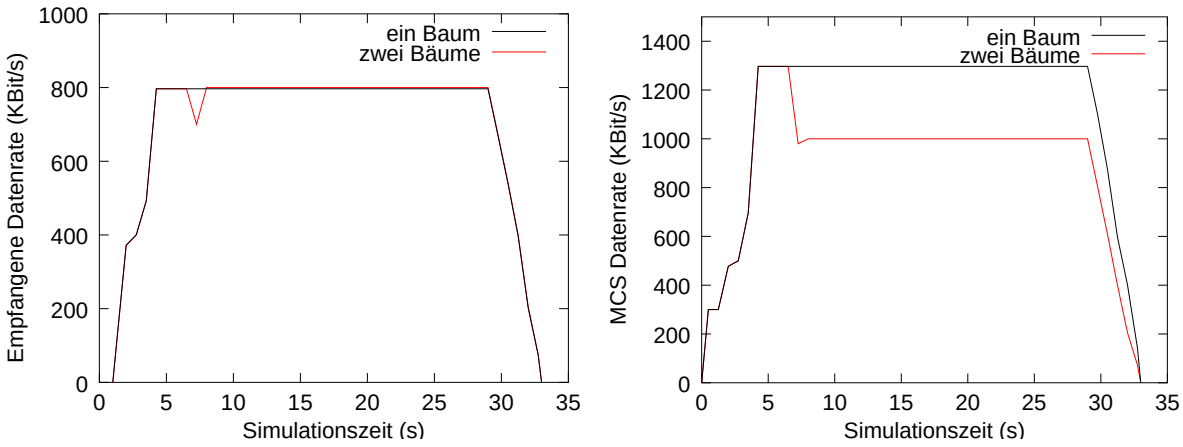


(b) Zwei Bäume.

Abbildung 6.46.: Beispiel-Szenarien für parallele Bäume.

Die Abbildung 6.47(a) zeigt die kumulierte empfangene Datenrate aller Empfänger für einen Baum und für zwei Bäume, die genau der gesendeten Datenrate von je 100 KBit/s pro Sender (800 KBit/s = 4 Sender * 100 KBit/s * 2 Empfänger) entspricht. Bei Zeitpunkt 6s wird der zweite Baum etabliert. Dies führt zu einer kurzen Unterbrechung im Datentransport, die die Ursache für den Knick in Abbildung 6.47(a) darstellt. Darauf wird weiter unten noch genauer eingegangen. Ansonsten sind beiden Kurven deckungsgleich, was bedeutet, dass das Verfahren mit mehreren Bäumen für eine Gruppe anwendbar ist. Die Stufen bei Anstieg und Abfall der Datenraten in Abbildung 6.47 basieren auf Rundungsungenauigkeiten bei der Berechnung.

Die Abbildung 6.47(b) zeigt die summierten Ausgangsdatenraten der MCS. Nach dem Aufbau des zweiten Baumes zum Zeitpunkt 6s sinkt die MCS Datenrate um ca.



(a) Empfangene Datenrate.

(b) MCS Datenrate.

Abbildung 6.47.: Datenraten der Empfänger und MCS mit einem Baum und mit zwei Bäumen.

20%. Das ist dadurch zu erklären, dass weniger Server beteiligt sind, z. B. werden bei der Gruppenkommunikation mit einem Baum von Sender B5 Daten über MCS B2, MCS B1 und MCS A1 zu Empfänger A4 gesendet. Bei der Gruppenkommunikation mit mehreren Bäumen werden die Daten nicht über MCS B1 und MCS A1, sondern stattdessen über MCS A2 gesendet. D. h. die Daten werden über ein Zwischensystem weniger an Endsystem A4 geleitet. Hierdurch wird deutlich, dass die gesendeten (und auch die empfangenen) Datenraten der MCS bei der Verwendung mehrerer Bäume geringer sein muss als bei einem Baum, wenn lokal für jeden Baum ein MCS zur Verfügung steht. Bevor die Belastungen der MCS untersucht werden, soll jedoch die Unterbrechung in der Gruppenkommunikation während des Baumaufbaus untersucht werden.

Unterbrechungsdauer während des Baumaufbaus:
Die Dauer der Unterbrechung beim Aufbau eines neuen Baums entspricht vom Prinzip her der Unterbrechung bei der MCS-Ersetzung. Die Unterbrechung entsteht zwischen dem Verbindungsabbau der MCS, die für den neuen Baum bestimmt worden sind, und der vollständigen Etablierung des neuen Baums.

Für das obige Szenario in Abbildung 6.46 zeigt die Tabelle 6.6 die Zusammensetzung der Unterbrechungsdauer. Insgesamt dauert die Unterbrechung etwa 320ms. Die ermittelten Werte setzen sich wie bei der MCS-Ersetzung aus drei Komponenten zusammen. Die Signalisierung der Controller hat einen Anteil von 33,4%. Dieser Anteil ist deutlich höher als bei der Restrukturierung, und ist darin begründet, dass zuerst die Verbindungen abgebaut werden und der Aufbau der neuen Verbindungen erst nach den indirekten Aktualisierungen erfolgt. Der Anteil der ATM-Signalisierung beträgt 49,8% und ist wie auch bei der MCS-Ersetzung nicht zu vernachlässigen. Mit 16,8% wirkt sich noch die Verzögerung beim Datenaustausch aus.

Bei einer Hierarchie mit drei Ebenen ist die Dauer einer Unterbrechung ca. 380ms.

Zeitpunkt (ms) Signalisierung	Unterbrechung (ms) Datentransport	Beschreibung	Signalisierung
0	–	Höchster Controller sendet Aufbau-Nachricht	Controller
46,7	–	Lokale Controller empfangen Aufbau-Nachrichten	
47,9	0	Verbindungsabbau der MCS	
93,4	45,5	Aktualisierungen treffen beim höchsten Controller ein	
150,2	102,3	Aktualisierungen treffen bei den lokalen Controllern ein	
154,2	106,3	Verbindungsaufbau bei MCS A1 und B2	
265,6	217,7	Verbindungen werden akzeptiert	ATM
312,7	264,8	Verbindungen aufgebaut	
366,1	318,2	Verzögerung der Dateneinheiten	keine

Tabelle 6.6.: Zusammensetzung der Unterbrechungsdauer beim Aufbau eines neuen Baums.

Hiervon entfallen 167ms auf die Signalisierung. Diese Erhöhung sowohl bei der Gesamtdauer als auch beim Anteil der Controller-Signalisierung ergibt sich dadurch, dass mehr Controller aktualisiert werden müssen, da mehr Hierarchieebenen vorhanden sind. D. h. bei einem Szenario mit mehr Hierarchieebenen ist aufgrund der längeren Aktualisierungszeit die Dauer der Unterbrechung größer.

Beim Aufbau eines neuen Baums wird die Gruppenkommunikation nicht für die gesamte Gruppe, sondern nur für einen Teil der Gruppe unterbrochen. Betroffen sind hiervon die Daten aller Sender, die dem neu aufzubauenden Baum angegliedert werden. Diese Daten können während der Aufbauphase nicht mehr an die Empfänger weitergeleitet werden. Im obigen Beispiel (Abbildung 6.46) sind die Daten der Sender A5 und B5 betroffen.

Belastungsverteilung auf die MCS:

Das Ziel der Etablierung eines neuen Baums ist auch in einem ATM-Weitverkehrsnetz eine verbesserte Belastungsverteilung auf die MCS zu erreichen, wie es bei dem lokalen SkaGAN Ansatz (Kapitel 5) möglich ist. Zunächst soll die Belastungsverteilung an dem Szenario aus Abbildung 6.46 dargestellt werden. Dieses Szenario bietet insofern gute Voraussetzungen für die Belastungsverteilung, da hier ausreichend MCS für einen zweiten Baum vorhanden sind.

Die Tabelle 6.7 fasst die Ergebnisse der MCSBelastungen bei einem und bei zwei Bäumen zusammen. Zusätzlich ist noch die Änderung der Belastungen mit angegeben, die sich nach dem Wechsel von einem Baum auf zwei Bäume ergeben hat. Die Angaben in der Tabelle sind natürlich sehr spezifisch und lassen sich nicht verallgemeinern (was besonders für die prozentualen Angaben gilt), dennoch ist gut zu erkennen, dass die

MCS	Belastung bei		Änderung	
	einem Baum	zwei Bäumen	absolut	prozentual
A1	0,110	0,065	−45	−50%
A2	0,09	0,135	+45	+50%
B1	0,175	0,09	−85	−50%
B2	0,045	0,045	0	0%
Gesamt	0,42	0,335	−85	−21%

Tabelle 6.7.: Belastungen der MCS bei einem Baum und bei zwei Bäumen.

Gesamtbelastung der MCS verringert werden konnte. Dieses Ergebnis bestätigt auch die Daten aus Abbildung 6.47(b), die eine insgesamt verringerte Datenrate bei den MCS belegt haben.

Dennoch wird nicht bei allen MCS die Belastung verringert, bei MCS A2 erhöht sich sogar die Belastung. Zu erklären ist das dadurch, dass MCS A2 jetzt ein primärer MCS ist und zusätzlich die Daten von B2 weiterzuleiten hat (vgl. Abbildung 6.46). Dieses muss sogar so sein, damit die maximale Belastung gesenkt werden kann. Insgesamt betrachtet verteilt sich die Belastung besser bei zwei Bäumen als bei der Verwendung eines einzelnen Baums, auch wenn einzelne MCS dadurch stärker belastet werden.

Die folgenden Szenarien sollen die Auswirkungen der Verteilung der MCS auf die Belastungen zeigen. Dazu wird im obigen Beispiel (Abbildung 6.46) die Anzahl der MCS in den lokalen Netzen variiert. Für folgende Variationen sind die Belastungen der MCS bei ebenfalls unterschiedlichen Baumanzahlen ermittelt worden:

Netz A	2	2	2	2	1	3	3
Netz B	2	2	1	3	4	3	3
Anzahl Bäume	1	2	2	2	2	3	4

Bei den Kombinationen variiert das Verhältnis zwischen der Anzahl der MCS und der Anzahl der Bäume in den lokalen Netzen. Das führt zum einen dazu, dass MCS nicht in einen Baum einbezogen werden und zum anderen, dass mehrere Bäume einen MCS mehrfach nutzen. Diese Fälle sind aber gerade von einem besonderem Interesse, da die Annahme, dass immer genügend MCS für einen Baum vorhanden sind, nicht realistisch ist. Die Ergebnisse der Messungen sind in Tabelle 6.8 zusammengefasst. Die Tabelle 6.9 stellt die gemessenen Belastungen prozentual zur Gesamtbelastung dar. Die prozentualen Belastungen sind daher nur zeilenweise zu interpretieren, da die Gesamtbelastungen verschieden sind.

Bei der prozentualen Verteilung ist zu erkennen, dass der prozentuale Anteil in einem MCS sehr hoch ist, wenn er der einzige in einem lokalen Netz ist (z. B. bei 2*A, 1*B und 2 Bäumen in Tabelle 6.9). Ansonsten verteilen sich die Belastungen wie erwartet eher gleichmäßig auf die MCS. Interessant ist noch die Bewertung der Gesamtbelastung. Aus der Tabelle 6.8 ist deutlich ein Zusammenhang mit einer steigenden Anzahl Bäume und einer abnehmenden Gesamtbelastung abzulesen. Die Ursache liegt in der verringerten Anzahl primärer MCS. Ein primärer MCS muss die empfangenen Daten immer

Schema (# MCS)	Anzahl Bäume	MCS-Belastungen (in 1000stel) A1	A2	A3	B1	B2	B3	B4	Gesamt
2*A, 2*B	1	110	90	–	175	45	–	-	420
2*A, 2*B	2	65	135	–	90	45	–	-	335
2*A, 1*B	2	90	90	–	175	–	–	-	355
2*A, 3*B	2	65	135	–	90	0	45	-	335
1*A, 4*B	2	–	265	–	175	45	0	0	485
3*A, 3*B	3	65	0	65	90	0	45	-	265
3*A, 3*B	4	65	0	65	90	0	45	-	265

Tabelle 6.8.: Belastungen bei verschiedenen Kombinationen der MCS in den lokalen Netzen.

Schema (# MCS)	Anzahl Bäume	MCS-Belastung (in %) A1	A2	A3	B1	B2	B3	B4
2*A, 2*B	1	26,2	21,4	–	41,7	10,7	–	–
2*A, 2*B	2	19,4	40,3	–	26,8	13,4	–	–
2*A, 1*B	2	25,4	25,4	–	49,3	–	–	–
2*A, 3*B	2	19,4	40,3	–	26,8	–	13,4	–
1*A, 4*B	2	–	54,6	–	36,1	9,3	0	0
3*A, 3*B	3	24,5	0	24,5	34,0	0	17,0	–
3*A, 3*B	4	24,5	0	24,5	34,0	0	17,0	–

Tabelle 6.9.: Prozentuale Belastungen der MCS bei verschiedenen Kombinationen

noch zusätzlich auf einer separaten Verbindung an den nächst höheren MCS weiterleiten, was seine Belastung erhöht. Das Ergebnis, dass mehr Bäume die Gesamtbelastung verringern, ist daher nicht immer vollkommen zutreffend. Die Gesamtbelastung wird nur verringert, wenn dabei auch die Anzahl der Ebenen eines Baumes vermindert wird. Dies kann aber nicht immer in jedem Fall bei der Etablierung eines neuen Baumes garantiert werden.

Eine Ausnahme in Tabelle 6.8 bildet die Zeile 5 (1*A, 4*B). Hier ist die Belastung gegenüber den anderen gemessenen Belastungen maximal. Die Ursache hierfür ist, dass im lokalen Netz A nur ein einzelner MCS für zwei Bäume vorhanden ist (siehe auch den Abschnitt 'Weniger lokale MCS als Bäume' auf Seite 126 in Unterkapitel 6.3.2). Dieser MCS A2 muss beide Bäume bedienen und weist daher eine deutlich höhere Belastung auf.

Anhand der Messergebnisse ist zu erkennen, dass es für eine gute Belastungsverteilung auf die MCS entscheidend ist, dass in jedem lokalen Netz ein MCS pro Baum vorhanden ist. Unter dieser Voraussetzung kann die Belastung gut zwischen den vorhandenen MCS aufgeteilt werden. Ist ein Ungleichgewicht vorhanden, kann es sich sogar negativ auswirken, wenn ein weiterer Baum aufgebaut wird. Außerdem ist entscheidend, ob die Sender gleichmäßig auf die Bäume verteilt werden. Sind die Sender nicht gleich-

mäßig auf die Bäume verteilt, so ist die Belastung in dem einen Baum stets höher als im anderen.

6.5. Zusammenfassung

Dieses Unterkapitel behandelte eine Lösung für eine Gruppenkommunikationsunterstützung in ATM-Weitverkehrsnetzen. Zusätzlich sind eine Reihe von Messungen durchgeführt und bewerten worden, die es erlauben sollen, den globalen Ansatz von SkaGAN in seiner Leistungsfähigkeit einzuschätzen. Die Bewertung teilt sich hierzu in die Verwaltung und den Datentransfer auf. Der Datentransfer ist noch einmal unterteilt in das Basisschema und Erweiterungen. Alle drei Teile konnten weitestgehend getrennt voneinander analysiert werden.

Gruppenverwaltung

Bei der Gruppenverwaltung ist untersucht worden, inwieweit der hierarchische Ansatz bezüglich Gruppengröße, Netzwerkgröße und Gruppenanzahl skaliert. Hier lieferten die Messungen sehr befriedigende Ergebnisse. Bei der Verwaltung hängt die Anzahl der benötigten Nachrichten pro Gruppe von der Höhe Verwaltungsbaumes ab und ist unabhängig von der Gruppengröße. Die Einflüsse der Netzgröße spielen sich hauptsächlich in der benötigten Anzahl der Controller wieder. Hat ein Controller zu viele Endsysteme zu verwalten, so kann es hier schnell zu einem Engpass kommen. Dieses Verhalten ist analog zum MARS Konzept. Bei den hier durchgeführten Messungen zeigte sich ein Verhältnis von Controllern zu Endsystemen in der Größenordnung von 1 : 100 als angemessen. Dieses Ergebnis ist allerdings aus einer Simulation heraus entstanden, wo viele Parameter bzgl. des Aufwands bei der Nachrichtenverarbeitung vereinfacht worden sind, und kann somit nicht als verbindlich angesehen werden.

Im Verhältnis zur Gruppenanzahl wächst der Verwaltungs- und Signalisierungsaufwand linear. Das ist auch nicht weiter erstaunlich, da keinerlei Möglichkeit besteht, Gruppen zusammenzufassen. Die Anzahl der Daten, die pro Gruppe in einem Controller gespeichert werden müssen, ist aber verhältnismäßig gering, so dass hierdurch in heutigen Systemen keine Probleme oder Engpässe zu erwarten sind.

Datentransfer

Ein wichtiges Kriterium für den Datentransfer ist die Verzögerung zwischen der Gruppenanmeldung und der Ankunft des ersten Datenpaketes dieser Gruppe. Diese Verzögerung kann insbesondere bei einem Weitverkehrsnetz einen längeren Zeitraum einnehmen. Hier konnte gezeigt werden, dass sich die Anmeldeverzögerung mit Hilfe der Baumstruktur beim Datentransfer signifikant reduzieren lässt. Die Ursache liegt in der lokalen Begrenzung bei der Anmeldung eines Teilnehmers. Ist schon ein Teilnehmer in der Nähe bei der Gruppe angemeldet, so muss die Anmeldung der weiteren Teilnehmer nur noch lokal durchgeführt werden.

Die Belastungen der MCS werden beim Basisschema für den Datentransfer nicht beachtet und können somit auch nicht direkt beeinflusst werden. Die einzige Möglichkeit besteht in der Zuordnung verschiedener Gruppen auf unterschiedliche MCS. Diese Zuordnung ist aber zufällig und hängt von der Reihenfolge ab, in der sich die einzelnen Teilnehmer für die verschiedenen Gruppen anmelden. Wird von einer (geografisch) zufälligen Verteilung der Gruppenteilnehmer ausgegangen, ergibt sich eine akzeptable Verteilung der MCS-Belastungen. Wenn die Gruppen allerdings in ihrem Anmeldeverhalten korrelieren, ergibt sich auch eine Konzentration bei der MCS-Belastung.

Ein anderer Punkt, der untersucht worden ist, ist die Menge der Daten, die im Netz verschickt werden müssen, damit alle Teilnehmer die Daten erhalten können. Hier zeigt sich, dass das VC-Mesh-Schema in der Gesamtheit aller Daten immer die geringsten Datenmengen erzeugt. In starke Gegensatz steht hierzu beim VC-Mesh-Schema allerdings die Skalierbarkeit der Gruppenverwaltung. Das MCS-Schema erzeugt immer eine höhere Datenmenge als das VC-Mesh-Schema. Für das MCS-Schema ist interessant, dass die erzeugte Datenmenge annähernd unabhängig von der Position des MCS im ATM-Netz ist. Beim Ansatz von SkaGAN ist die generierte Datenmenge im Netz erwartungsgemäß um einiges höher als beim VC-Mesh- und MCS-Schema. Das hat als Ursache die Kommunikation zwischen den MCS und die nicht immer optimale (geografische) Wahl eines MCS.

Erweiterungen für den Datentransfer

Für eine verbesserte Lastverteilung sind bei SkaGAN zwei Verfahren, die MCS-Ersetzung und parallele Bäume, vorgestellt worden, die in Weitverkehrsnetzen eine Lastverteilung bei der Gruppenkommunikation ermöglichen.

Durch die MCS-Ersetzung ist es möglich, einen aktiven primären MCS durch einen Ersatz-MCS auszutauschen. Die Messungen haben gezeigt, dass hierdurch die Ende-zu-Ende-Verzögerung und die Gesamtbelastung der MCS gesenkt werden konnte. Das ist vor allem dadurch möglich, dass MCS in höheren Ebenen ausgetauscht werden können, die sich nicht mehr an zentraler Position im Netz in Bezug auf die Gruppe befinden. Die Häufigkeit der MCS-Ersetzungen ist ebenfalls von Belang, da hierbei Unterbrechungen in der Gruppenkommunikation entstehen können. Die Häufigkeit der Unterbrechungen hängt dabei mit der Baumstruktur der Gruppenverwaltung zusammen. Je mehr Ebenen vorhanden sind, desto mehr MCS-Ersetzungen finden statt, aber haben auf die Gruppe meist nur eine lokale Auswirkung.

Durch die Nutzung von parallelen Bäumen ist es ebenfalls möglich, die Gesamtbelastung der MCS zu senken. Bei mehreren Bäumen ist die durchschnittliche Baumhöhe der Datentransferbäume geringer als bei einem Baum, wodurch sich die Anzahl der Zwischensysteme und somit die MCS-Belastung verringert. Dieser Effekt stellt sich jedoch nur ein, wenn ausreichend MCS in den lokalen Teilnetzen vorhanden sind.

Insgesamt konnte gezeigt werden, dass eine skalierbare Gruppenkommunikationsunterstützung für ATM-Netze möglich ist. Der globale Ansatz von SkaGAN basiert auf einer Baumstruktur und ist hierdurch in der Lage, große Gruppen in einem Weitverkehrsnetz zu unterstützen.

7. Zusammenfassung und Ausblick

7.1. Ergebnisse der Arbeit

In der hier vorliegenden Arbeit ist ein Konzept (SkaGAN) entwickelt worden, das eine rechnergestützte Gruppenkommunikation über ATM-Netzen ermöglicht. Einen Überblick Über die von SkaGAN erfüllten Kriterien gibt Tabelle 7.1 (zu den einzelnen Kriterien siehe Kapitel 3, Seite 29). Das Konzept ist ein Lösungsvorschlag für den Bereich der lokalen ATM-Netze und für den Bereich der ATM-Weitverkehrsnetze (global). Der Schwerpunkt liegt dabei auf der Berücksichtigung der Skalierbarkeit, besonders im Backbone-Bereich, bei ATM-Weitverkehrsnetzen. Die ATM-Technologie ist zwar in der Lage, multimediale Anwendungen mit hohen Anforderungen zu bedienen, bietet aber keine effiziente Unterstützung für die Gruppenkommunikation an.

	SkaGAN	
Kriterium	lokal	global
Datentransport		
Schema	mehrere MCS	Baum
Verkehrskonzentration	verteilt auf MCS	
Verzögerung	moderat	moderat-erhöht
Datenformat	Einkapselung	
Dienstgüteunterstützung	nein	nein
Fehlertoleranz	hoch	hoch
Ressourcenbedarf	$2N + L(N+1)$	$2N + N\frac{k+1}{k} + 2\log_k N$
Verwaltung		
Organisation	zentral	verteilt, aggregiert
Ausfallsicherheit	SPoF	Teilbaumausfall
Signalisierungsaufwand	MARS + L * MCS	$\log_k N$
IDMR-Protokolle	nein	nein

Tabelle 7.1.: Bewertung des SkaGAN-Konzeptes.

In der Literatur sind eine Reihe von Lösungsansätzen für die Gruppenkommunikationsunterstützung vorgeschlagen worden. Die meisten Ansätze behandeln aber nur den Bereich der lokalen ATM-Netze und skalieren nicht für eine größere Anzahl Teilnehmer oder bei größeren Netztopologien. Die Lösungsansätze sind begrenzt, da alle

entweder auf dem VC-Mesh- oder dem MCS-Schema aufbauen, welche beide gravierende Nachteile bei einer steigenden Anzahl von Teilnehmern aufweisen. Andere Ansätze propagieren Lösungen innerhalb der ATM-Schicht, wozu aber Änderungen bei allen ATM-Komponenten notwendig wären. Daher ist der Einsatz dieser Verfahren nur schwer und in einem langwierigen Prozess durchsetzbar. Die Lösungsansätze im Bereich der Weitverkehrsnetze behandeln meist nur die Gruppenverwaltung oder die Interaktion mit anderen Multicast-Routingprotokollen und verwenden für den Datentransfer dieselben Schema wie bei lokalen Netzen. Diese Verfahren stellen daher auch keine adäquaten Lösungen dar.

Für die Lösung einer skalierbaren Gruppenkommunikationsunterstützung ist in dieser Arbeit zunächst der MARS-Ansatz zugrunde gelegt worden. Der MARS ist ein von der IETF standardisiertes Konzept zur IP-Multicast-Emulation in lokalen ATM-Netzen. Wenn die Anzahl der Gruppenteilnehmer begrenzt ist, stellt der MARS eine akzeptable Lösung dar. Bei einer größeren Anzahl (>200 nach [64]) lokaler Teilnehmer zeigt der MARS-Ansatz aber gravierende Nachteile. Daher ist in dieser Arbeit ein Ansatz entwickelt worden, der im Bereich der lokalen ATM-Netze eine bessere Skalierbarkeit in der Gruppenkommunikation ermöglicht.

Im Bereich der Gruppenkommunikationsunterstützung für lokale Netze ist das MCS-Schema mit mehreren MCS erweitert worden. Gegenüber Lösungsvorschlägen aus der Literatur wird hier eine gezielte Lastverteilung auf die vorhandenen MCS durchgeführt. Hierdurch können Datenkonzentrationen in den MCS vermieden werden, solange die Datenmenge nicht die Gesamtkapazität der MCS überschreitet. Die Ende-zu-Ende-Verzögerung der Datenpakete wird ebenfalls reduziert, da die MCS gleichmäßig belastet werden. Dieses ermöglicht auch den Einsatz von mehr sendenden Teilnehmern als beim MARS, ohne dass es zu Engpässen kommen muss. Der Einsatz mehrerer MCS erhöht zudem die Ausfallsicherheit, die Aufgaben eines MCS können auf andere MCS übertragen werden.

Im Weitverkehrsbereich ist der MARS-Ansatz ebenfalls nur sehr eingeschränkt einsetzbar. Hier kommt vor allem zusätzlich zu der Problematik des Datentransfers die Problematik der Gruppenverwaltung hinzu. Die Gruppenverwaltung sollte für Weitverkehrsnetze dezentral und verteilt organisiert sein und vor allem eine Möglichkeit bieten, Teilnehmer zu aggregieren. Ansonsten steigt der Signalisierungsaufwand bei größeren Gruppen sehr stark an.

Bei SkaGAN ist für die Gruppenverwaltung eine hierarchische Struktur eingesetzt worden. Mithilfe eines Baumes können die Teilnehmer in den Knoten aggregiert und abstrahiert werden. Zwischen den Knoten im Baum werden nur noch die Änderungen in der Gruppenmitgliedschaft weitergegeben. Dieses Verfahren skaliert sehr gut in Bezug auf die Gruppengröße und die Netzwerktopologie.

Die hierarchische Gruppenverwaltung stellt die Grundlage für den Datentransfer zwischen den Teilnehmern einer Gruppe dar. Für den Datentransfer ist dabei ebenfalls eine hierarchische Baumstruktur gewählt worden. Während die Teilnehmer lokal das MCS-Schema nutzen, werden die MCS global mit einer Baumstruktur untereinander verbunden. Die lokalen MCS übernehmen dabei auch Aufgaben in den höheren Baumknoten. Dieses Baumschema ist den Core Based Trees ähnlich und ein Baum wird für

jede Gruppe separat aufgebaut und verwaltet.

Bei der Baumstruktur ist versucht worden, die Anzahl der benötigten Zwischensysteme (MCS) gering zu halten. Konkret bedeutet dies, dass die Lokalität der Teilnehmer berücksichtigt wird. Die Anzahl der Zwischensysteme nimmt nur bei steigender Entfernung der Teilnehmer zu, lokale Teilnehmer werden nach Möglichkeit direkt verbunden. Hieraus resultiert auch ein weiterer Vorteil, die Verzögerung bei der Anmeldung kann im Mittel deutlich reduziert werden. Existiert bei der Anmeldung eines Teilnehmers bereits ein weiterer Teilnehmer in der ‚Nähe', so wird nur eine lokale Verbindung etabliert, was in wesentlich kürzerer Zeit erfolgen kann.

Insgesamt konnte für weit verteilte Gruppen die mittlere Verzögerung bei der Datenübertragung gegenüber dem MCS-Schema verringert werden, da die Lokalität der Teilnehmer bei SkaGAN Berücksichtigung findet. Hingegen ist die Gesamtdatenmenge im ATM-Netz bei SkaGAN erhöht. Dies ist durch die nicht immer optimale Wahl der MCS im ATM-Netz und die zusätzlich hinzugekommene Kommunikation zwischen den MCS im Baum zu erklären. Dennoch kann insgesamt mit der hierarchischen Gruppenverwaltung eine gute Skalierbarkeit der Gruppenkommunikation in ATM-Weitverkehrsnetzen erreicht werden.

Zusätzlich zu diesem Basisschema sind mehrere Erweiterungen entwickelt worden, die die Behandlung von dynamischen Gruppen verbessern und bei Gruppen mit erhöhten Datendurchsatz und vielen sendenden Teilnehmern eine bessere Skalierbarkeit ermöglichen.

Bei dynamischen Gruppen ändern sich die Teilnehmer und damit kann sich die topologische Lage der Gruppe im Netz ebenfalls ändern. Das führt dazu, dass MCS im Netz eine ungünstige Position haben und die Netzwerkbelastung und Verzögerung erhöht wird. Um diesem Problem entgegen zu wirken ist ein Verfahren zur MCS-Ersetzung entwickelt worden, womit die Aufgaben eines MCS auf einen anderen MCS im Netz delegiert werden können. Die Ersetzung eines MCS hängt dabei nicht ausschließlich von seiner Lage im Netz ab, sondern auch von seiner momentanen Belastung. Ist ein MCS durch mehrere Gruppen sehr hoch belastet, können einzelne Gruppen auf andere MCS verteilt und somit die Belastung der MCS vermindert bzw. ausgeglichen werden.

Bei Gruppen mit vielen aktiven Sendern stellt sich hingegen das Problem, dass alle Datenpakete über einen Baum an alle Empfänger verteilt werden, was zu hohen Datenkonzentrationen und erhöhten Belastungen der involvierten MCS führt. Um diese Situation zu verbessern, ist eine Unterstützung für mehrere Bäume pro Gruppe eingeführt worden. Die Gruppe wird hierzu weiter in Untergruppen aufgeteilt, die jeweils einen eigenen Baum verwenden. Damit ist dieses Verfahren zwischen den Core Based Trees, die eine Gruppe über einen Spannbaum verbinden, und den Source Based Trees, die einen Spannbaum pro Sender etablieren, angesiedelt.

Ein weiterer Vorteil bei der Aufteilung in mehrere Bäume ist die Reduktion der benötigten Zwischensysteme. Da jeder Baum jetzt weniger sendende Teilnehmer hat, verringert sich die Anzahl der benötigten MCS und somit die Baumhöhe. Dadurch wird auch die Ende-zu-Ende-Verzögerung verringert. Der Einsatz mehrerer Bäume eignet sich besonders in Kombination mit mehreren lokalen MCS. Die Bäume können dann auf die vorhandenen lokalen MCS aufgeteilt werden, wodurch eine wesentlich bessere

Lastverteilung möglich wird.

Beide Erweiterungen, die MCS-Ersetzung und Gruppen mit mehreren Bäumen, können auch kombiniert eingesetzt werden. Dadurch wird die Behandlung von Gruppen sehr flexibel und zwei Kernprobleme bei der Gruppenkommunikation, die Gruppendynamik und die Gruppengröße, können adäquat behandelt werden. Im Ergebnis zeigt sich eine gute Skalierbarkeit bei der Gruppenkommunikation, sowohl in lokalen ATM-Netzen als auch in ATM-Weitverkehrsnetzen.

7.2. Ausblick

Das in dieser Arbeit vorgestellte Konzept für eine skalierbare Gruppenkommunikation in ATM-Netzen (SkaGAN) stellt einen realisierbaren Lösungsvorschlag dar. Für dessen Umsetzung vom Modell auf ein konkretes ATM-Netz sind aber noch eine Reihe von Randbedingungen zu beachten.

Zum einen ist eine bessere Integration in die Netzwerkschicht der Endsysteme oder Multicast-Router notwendig, damit IP-Multicast-Datenpakete transparent für die Anwendungen über ATM transportiert werden können. Die für die Simulationen verwendete Netzwerkschicht stellt nur ein Grundgerüst mit minimalem Funktionsumfang dar und der bisher für SkaGAN entwickelte Ansatz [48] erfordert geringfügige Änderungen und eine Neukompilierung der Multicast-Anwendungen.

Das erfordert insofern einen etwas höheren Aufwand, als dass bei IP Multicast die Sender Datenpakete an eine Gruppe schicken können, ohne dieser Gruppe anzugehören. Das ist bei SkaGAN nicht möglich, jeder Sender an eine Gruppe muss auch Teilnehmer in der Gruppe sein. Daher ist für die IP-Multicast-Emulation immer zuerst eine implizite Senderanmeldung durchzuführen. Eine Abmeldung des IP-Multicast-Senders muss ebenso implizit in der Emulationsschicht erfolgen.

Eine Anbindung an bestehende Multicast-Routingprotokolle ist bei SkaGAN ebenfalls noch nicht vorhanden. Hier sollte eine Schnittstelle geschaffen werden, die die Anbindung von IDMR-Protokollen ermöglicht. Hierzu kann aber auf bestehende Konzepte, wie z. B. IMSS (Unterkapitel 3.2.7, ab Seite 42) zurückgegriffen werden.

Bei der Entwicklung der Erweiterungen zur MCS-Ersetzung und bei mehreren Bäumen ist hauptsächlich auf die eigentlichen Verfahren eingegangen worden. Wann ein MCS ersetzt oder ein Baum in zwei Bäume aufgeteilt wird, ist nicht genauer untersucht worden. In dieser Arbeit sind diese Entscheidungen aufgrund von Schwellwertüberschreitungen durchgeführt worden, wobei diese Schwellwerte aber nicht genauer angegeben worden sind. Hier besteht weiterer Forschungsbedarf, zum einen inwieweit ein Entscheiden aufgrund von Schwellwerten angemessen ist, und zum anderen, welche konkreten Schwellwerte gewählt werden sollten, um die idealen Zeitpunkte für den Einsatz der Erweiterungen zu finden.

A. Netzwerksimulationswerkzeug OpNet

OpNet™ ist ein Netzwerksimulator und eine Software-Entwicklungsumgebung für Kommunikationsnetze, Protokolle und verteilte Systeme der Firma MIL3 [65]. In OpNet können Modelle konstruiert werden, um anschließend deren Systemverhalten und Leistung zu untersuchen. Für die Konstruktion der Modelle wird eine Modellierungsumgebung für den Entwickler bereitgestellt.

Typische Modelle sind beispielsweise Local-Area-Network- und Wide-Area-Network-Modelle zur Leistungsuntersuchung, Kommunikationsarchitekturen in Forschung und Entwicklung und mobile Funkübertragungsnetzwerke. Alle Modelle haben gemeinsam, dass sie hierarchisch strukturiert werden. Diese hierarchische Strukturierung spiegelt sich direkt im Aufbau von OpNet wieder, welches in Modellierungsbereiche (model domains) gegliedert ist.

Neben der hierarchischen Gliederung bietet OpNet Objektorientierung, Unterstützung durch eine grafische Oberfläche, High-Level-Programmiersprache und automatische Generierung von Simulationen und anwendungsspezifischen Statistiken. Die High-Level-Programmiersprache ist Proto-C, eine C/C++-Sprache mit OpNet-spezifischen Erweiterungen.

A.1. Modellimplementierung

Die Modellimplementierung läuft in mehreren Phasen ab. Diese Phasen sind die Modellspezifikation, Datensammlung und Simulation sowie die Analyse. Zusammen bilden diese drei Phasen einen (Projekt-) Zyklus (vgl. Abbildung A.1).

Begonnen wird mit der Spezifikation, auf welche die Datensammlung und Simulation mit abschließender Analyse folgen. Aufgrund von Unterschieden zwischen den Simulationsergebnissen und den Zielen kann sich ein erneuter Zyklus, beginnend mit einer Neuspezifikation, anschließen, womit der Zyklus geschlossen ist.

Die Spezifikation der Modelle in OpNet wird hierarchisch strukturiert. Die drei sich daraus ergebenden Funktionsbereiche (model domains) entsprechen in ihrer Art realen Netzwerken.

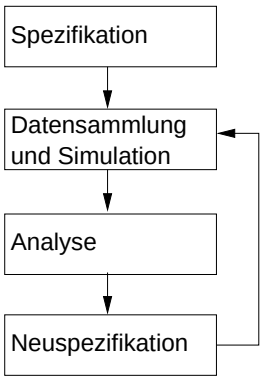


Abbildung A.1.: Projektzyklus in OpNet.

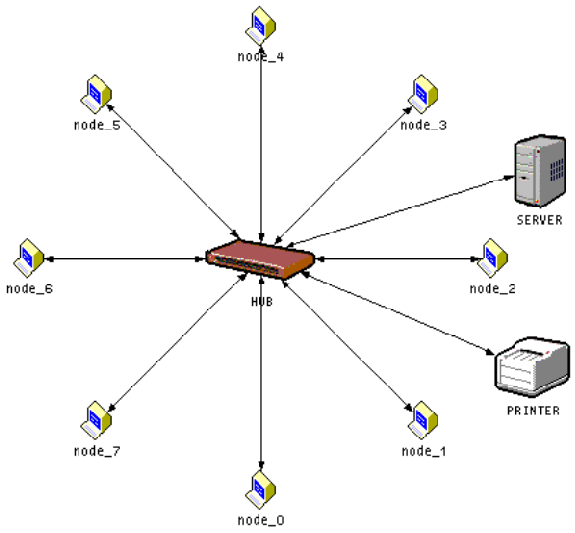


Abbildung A.2.: Beispielnetzwerk.

A.1.1. Modellierungsbereiche

Netzwerkmodell

Der oberste Modellierungsbereich ist das Netzwerk (network domain). Erstellt wird ein Netzwerk mit dem Projekteditor. Es wird durch Subnetze, Knoten und Verbindungen repräsentiert.

Knoten sind über Verbindungen kommunizierende Einheiten. Beispiele für Knoten sind Router, Brücken, Endsysteme, aber auch Satelliten. Jedem instanziierten Knoten im Netzwerk kann ein entsprechendes Verhalten über eine Knoteninstanz zugeordnet werden. Die Knoteninstanz wird mit dem Knoteneditor definiert.

Die Verbindungen im Netzwerk sind Punkt-zu-Punkt-Verbindungen (simplex oder duplex), Busse oder Funkverbindungen. Ferner können die Netzwerke hierarchisch strukturiert werden. Hierzu können Subnetze definiert werden, die wiederum aus Subnetzen, Knoten und Verbindungen bestehen.

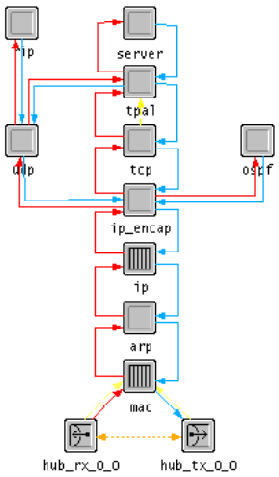


Abbildung A.3.: Knotenmodell in OpNet.

Ein Beispielnetzwerk zeigt die Abbildung A.2. Abgebildet ist ein sternförmig angeordnetes Netzwerk mit einem zentralen Hub und einem Server. Verbunden sind die Endsysteme und der Server mit dem Hub über Punkt-zu-Punkt-Verbindungen.

Knotenmodell

Die Endsysteme im Knotenbereich (node domain) können mit dem Knoteneditor definiert und verändert werden. Knotenmodelle bestehen aus Modulen und Verbindungen (connections). Die Module werden in zwei Gruppen unterteilt. In der ersten Gruppe sind Nachrichtengeneratoren, verschiedene Transmitter und Empfänger und in der zweiten Gruppe Warteschlangen und Prozessoren. Der Unterschied zwischen den beiden Modulgruppen ist, dass in der ersten Gruppe vorgegebene Module sind und in der zweiten weitgehend frei programmierbare Module.

Für das Verbinden der Module im Knotenmodell gibt es drei Typen von Verbindungen. Der erste Typ sind die Packet-Streams, die formatierte Nachrichten (Pakete) zwischen den Modulen befördern. Der zweite Typ sind Statistic Wires, die numerische Werte oder Kontroll-Informationen transportieren, und der dritte und letzte Typ sind die Logical Associations, die Verbindungen zwischen Modulen anzeigen. Sie werden speziell zwischen Sendern und Empfängern verwendet. Ein Beispiel hierzu zeigt die Abbildung A.3.

Die Abbildung A.3 zeigt das Knotenmodell eines Ethernet Servers, welches aus verschiedenen Modulen und Verbindungen besteht. Die dargestellten Verbindungen sind Statistic Wires und Packet Streams.

Prozessmodell

Das Verhalten von programmierbaren Modulen kann im Prozessbereich mit dem Prozesseditor verfeinert werden. Ein Prozess ist eine Instanz eines Prozessmodells und operiert

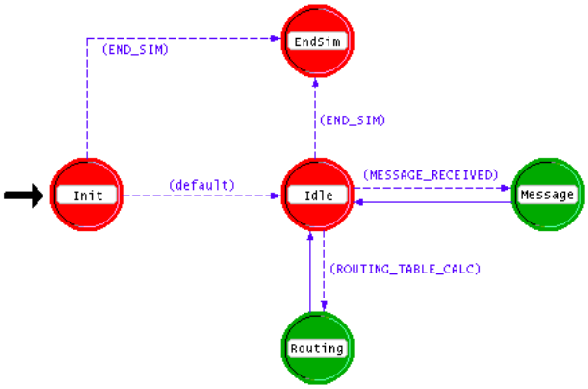


Abbildung A.4.: Prozessmodell in OPNET.

in einem Modul. Zu Beginn existiert genau ein Prozess pro Modul. Dieser Anfangsprozess kann weitere Kindprozesse dynamisch erzeugen.

Prozesse reagieren auf Interrupts (Unterbrechungen), welche anzeigen, dass wichtige Ereignisse wie der Ablauf eines Timers (Zeitgebers) oder das Eintreffen einer Nachricht eintreten. Die Interrupts in OpNet stellen das Pendant zu Ereignissen dar. OpNet ist ein Ereignis-gesteuerter Simulator. Wird ein Prozess unterbrochen, schließen sich Aktionen als Antwort darauf an und der Prozess wird angehalten, bis ein neuer Interrupt erfolgt.

Prozessmodelle werden in Proto-C verfasst, welches aus grafischen Zustandsübergangsdiagrammen, aus Teilen der Programmiersprache C/C++ und einer Bibliothek von Prozeduren aus dem Simulationskern besteht. Ein Zustandsübergangsdiagrammen zeigt die Abbildung A.4. Dargestellt ist das Modul `ospf` aus Abbildung A.3, welches den Open-Shortest-Path-First-Algorithmus implementiert.

Es gibt im Prozessmodell zwei verschiedene Arten von Zuständen, forced und unforced. Der Unterschied zwischen beiden Zuständen ist, dass unforced-Zustände ihren Prozess nach dem Eingangsanweisungsblock stoppen. In der Abbildung A.4 ist beispielsweise der `Init`-Zustand ein unforced-Zustand, der normalerweise zu Beginn der Simulation aufgerufen wird. In ihm werden gewöhnlich Prozessvariablen initialisiert. Prozessvariablen speichern Werte, die in allen Zuständen abgefragt und geändert werden können.

Verlassen werden kann der `Init`-Zustand über die `default`- oder die `END_SIM`-Transition. Beides sind bedingte Transitionen (conditional Transitions). Die `default`-Transition wird ausgeführt, wenn ein Interrupt ausgelöst wird und keine andere Bedingung zutreffend ist. Im Gegensatz dazu wird die `END_SIM`-Transition ausgeführt, wenn die Bedingung `END_SIM` zutreffend ist.

Die Zustände `Message` und `Routing` sind forced-Zustände. Forced-Zustände sind nicht blockierende Zustände. D. h. sie werden nach Eintritt und Abarbeitung wieder über die unbedingte Transition (unconditional Transition) verlassen. Zustände bestehen aus einem Eingangs- und einem Ausgangsanweisungsblock. Diese Anweisungsblöcke führen einen Programmcode bei Eintritt bzw. bei Verlassen des Zustands aus. Angegeben wird dieser Code in der Programmiersprache C.

Zur Unterstützung der drei Modellierungsbereiche gibt es einige weitere graphische

Editoren wie z. B. den Paketformateditor oder den ICI Editor (Interface Control Information Editor). Diese werden verwendet, um spezielle Schnittstellen und Abhängigkeiten zwischen den Modellen spezifizieren zu können.

B. OpNet-Prozessmodelle von SkaGAN

Das in den Kapiteln 4, 5 und 6 vorgestellte Modell zur Gruppenkommunikationsunterstützung in ATM-Netzen (SkaGAN) ist als Simulationsmodell für den Netzwerksimulator OpNet (Anhang A) realisiert worden. Das Simulationsmodell orientiert sich dabei genau an dem entwickelten Konzept. Die drei Komponenten Endsystem, MCS und Controller und allgemeine Module (Nachrichtentransport, Verbindungsmanagement) sind umgesetzt worden. Da das Simulationswerkzeug OpNet nur ATM-Punkt-zu-Punkt-Verbindungen unterstützt, ist das ATM-Modell um Punkt-zu-Mehrpunkt-Verbindungen ergänzt worden. Die einzelnen Teile des Simulationsmodells werden im Folgenden vorgestellt.

B.1. ATM-Punkt-zu-Mehrpunkt-Verbindungen

Für die Erweiterung von Punkt-zu-Punkt- auf Punkt-zu-Mehrpunkt-Verbindungen sind umfangreiche Änderungen und Erweiterungen im gesamten ATM-Modell von OpNet notwendig gewesen:

- Bei der Weiterleitung von ATM-Zellen ist eine Duplizierung der Zellen hinzugekommen, um die Zellen an mehrere Ausgangs-Ports verteilen zu können.
- Die Signalisierung in den Endsystemen und den Schalteinheiten ist entsprechend den Vorgaben von UNI3.1 erweitert worden. Nachdem eine Punkt-zu-Punkt-Verbindung aufgebaut ist, können dieser Verbindung Teilnehmer hinzugefügt und entfernt werden.
- Für AAL5 wird in der AAL-Schicht das Protokoll SSCOP (Service Specific Connection Oriented Protocol) eingesetzt. Hier musste eine Erweiterung stattfinden, um eine Betrieb mit mehreren Endpunkten zu ermöglichen.
- Die Schnittstelle zur Anwendungsschicht ist um weitere Kommunikations-Primitive erweitert worden, damit mehrere Teilnehmer pro Verbindung gehandhabt werden können.

Die Abbildung B.1 zeigt eine ATM-Schalteinheit mit vier Ports, ein Port besteht aus jeweils einem Eingang und einem Ausgang. Ankommende Zellen werden im Prozess

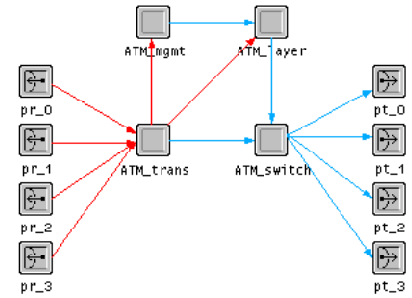


Abbildung B.1.: Knotenmodell einer ATM-Schalteinheit mit 4 Ports.

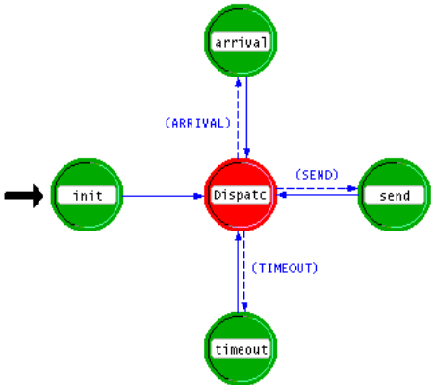


Abbildung B.2.: Prozessmodell für einen zuverlässigen Nachrichtentransport.

ATM_trans untersucht, ob sie Signalisierungs-, Daten- oder Kontrollzellen sind. Je nach Typ werden sie entweder direkt an die ATM-Schalteinheit (ATM_switch), die Verwaltung mit der Signalisierung (ATM_mgmt) oder an die AAL-Schnittstelle (ATM_layer) weitergeleitet.

B.2. Allgemeine Module

Das Prozessmodell für den zuverlässigen Nachrichtentransport (siehe auch Unterkapitel 4.2.2 ab Seite 61) zeigt Abbildung B.2. Ankommende Datenpakete werden im arrival-Zustand bearbeitet (Bestätigung zurücksenden und Datenpakete weitergeben). Die Zustände send und timeout gehören zusammen. Der Zustand send verschickt die Datenpakete und vergibt an diese entsprechende Referenznummern und der Zustand timeout sorgt für eine Übertragungswiederholung, wenn keine Bestätigung für das versendete Datenpaket eingetroffen ist. Dieser Prozess kommuniziert mit seinem Vaterprozess, und das ist entweder ein Endsystem, ein MCS oder ein Controller.

Das Management von ATM-Verbindungen (siehe auch Unterkapitel 4.2.3 ab Seite 63) geschieht in dem in Abbildung B.3 dargestellten Prozess. Es gibt pro ATM-Verbindung vier mögliche Aktionen (Setup, Release, AddParty und DropParty), die an die ATM/AAL-Schnittstelle weitergegeben werden. Jede Aktion muss von der AAL-

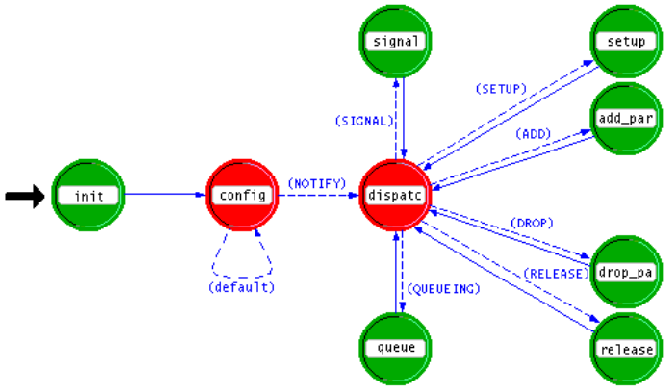


Abbildung B.3.: Prozessmodell für das Verbindungsmanagement.

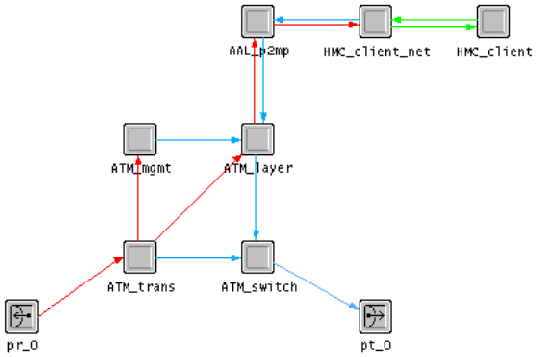


Abbildung B.4.: Knotenmodell des Endsystems.

Schicht positiv oder negativ bestätigt werden. Solange dies nicht geschehen ist, werden die Aktionen in einer Warteschlange (Zustand queue) zwischengespeichert. Für die Annahme von Verbindungen ist der Zustand signal zuständig. Hier wird jede ankommende Verbindung angenommen und der Vaterprozess benachrichtigt.

B.3. Endsystem

Das in Abbildung B.4 als Knotenmodell dargestellte Endsystem hat dieselben ATM-Prozesse wie eine ATM-Schalteinheit (Abbildung B.1), nur das jetzt AAL-Datenpakete über den Prozess AAL_p2mp verarbeitet werden können. Damit ist auch eine Schnittstelle zur Anwendungsschicht vorhanden. Die Anwendungsschicht des Endsystems besteht bei SkaGAN aus einem Prozess (hmc_client_net) für die Netzwerkschicht und einem Prozess (hmc_client) für die Emulation der Gruppenkommunikationsanwendungen.

Der Prozess zur Emulation der Anwendungen kann mehrere Anwendungen gleichzeitig nachbilden. Er initiiert Gruppenbei- und -austritte und ist für das Senden und Empfangen von Datenpaketen verantwortlich. Für das Senden können verschiedene Verteilungen (Exponentialverteilung, eine Heavy-Tailed-Verteilung oder keine Verteilung) für die Zwischenankunftszeiten und die Paketgrößen ausgewählt werden.

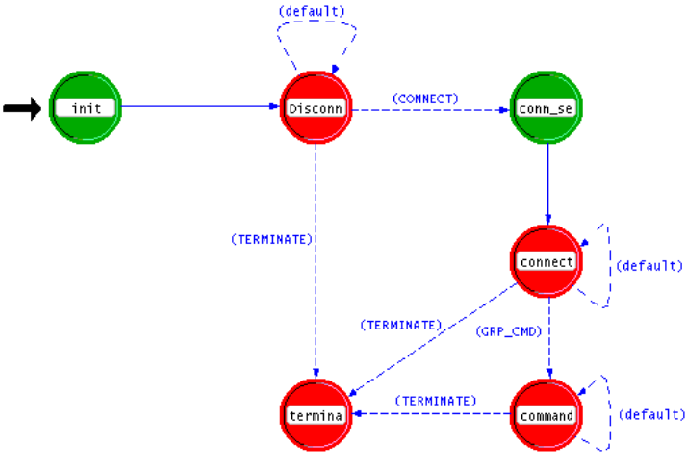


Abbildung B.5.: Prozessmodell der Netzwerkschicht des Endsystems.

Der Prozess für die Netzwerkschicht ist in Abbildung B.5 dargestellt. Zuerst wird eine Verbindung zum nächsten Controller aufgebaut (Zustand conn_setup), erst hiernach kann das Endsystem an der Grupenkommunikation teilnehmen. Die weiteren Aufgaben der Netzwerkschicht sind die Weitergabe von Gruppenbei- und -austritten, wobei die entsprechenden Nachrichtenformate generiert werden müssen. Die Datenpakete der Anwendungen müssen den zugehörigen ATM-Verbindungen übergeben werden und umgekehrt. Der Auf- und Abbau von ATM-Verbindungen zum MCS ist nach Anweisungen des Controllers durchzuführen (Zustand command).

B.4. MCS

Der MCS (Unterkapitel 5.3 ab Seite 67) hat dieselben ATM/AAL-Prozesse wie das Endsystem. Die eigentliche Funktionalität des MCS ist in einem Prozess (hmc_server) untergebracht, wie Abbildung B.6 zeigt. Dies ermöglicht auch eine Integration des MCS in eine ATM-Schalteinheit. Das OpNet-Modell des MCS-Prozesses ist in Abbildung B.7 dargestellt. Im Kern besteht der Prozess aus zwei Teilen, der Verwaltung und der Datenpaketverarbeitung.

Die Datenpaketverarbeitung besteht aus vier Zuständen, die das Verhalten des MCS nachbilden. Der Zustand incoming nimmt alle Pakete von der AAL-Schicht entgegen und speichert diese in der CopyIn-Warteschlange. Der Zustand input stellt den CopyIn-Bearbeitungsprozess dar. Anschließend kommen die Datenpakete in die Lookup-Warteschlange und der Lookup-Bearbeitungsprozess (Zustand lookup) verifiziert den Datenpaketkopf anhand der eigenen Datenstrukturen. Ist die zugehörige ATM-Verbindung für dieses Datenpaket gefunden, so wird es in der CopyOut-Warteschlange abgelegt. Der Zustand output übergibt das Datenpaket wieder an die AAL-Schicht.

Die Verwaltung des MCS besteht aus drei Zuständen (load, update, timer). Im Zustand update werden ankommende Signalisierungsnachrichten vom Controller angenommen und die Datenstrukturen entsprechend angepasst. Hierzu gehört auch der Auf-

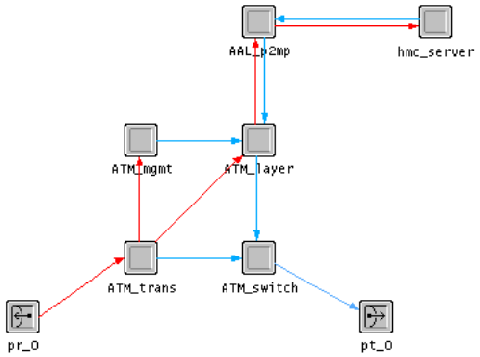


Abbildung B.6.: Knotenmodell des MCS.

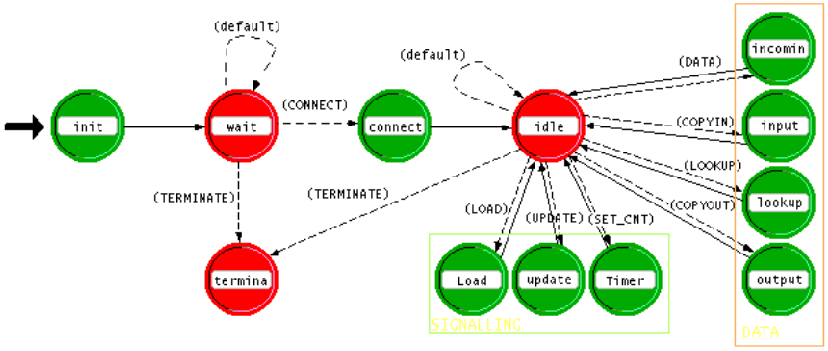


Abbildung B.7.: Prozessmodell des MCS.

und Abbau von entsprechenden ATM-Verbindungen. Belastungsanfragen vom Controller werden im Zustand load bearbeitet und umgehend beantwortet. Im Zustand timer wird sekündlich die Belastung des MCS gemessen und ausgewertet.

B.5. Controller

Das Knotenmodell des Controllers zeigt Abbildung B.8 (siehe auch Unterkapitel 4.1 ab Seite 4.1 und Unterkapitel 6.2.3 ab Seite 105). Die ATM-Prozesse sind identisch mit dem MCS und dem Endsystem, die Verbindung zur ATM/AAL-Schicht übernimmt der Prozess HMC_ctrl (Abbildung B.9). Die eigentliche Kontrolllogik des Controllers ist in zwei Prozesse aufgeteilt: local ist für die lokalen Teilnehmer und MCS zuständig (Abbildung B.10) und global für die Signalisierung in Weitverkehrsnetzen (Abbildung B.11).

Der in Abbildung B.9 dargestellte Prozess HMC_ctrl ist für die Signalisierungsverbindungen zuständig. Hierzu gehört die Annahme von Verbindungen von lokalen Systemen (Endsysteme und MCS) und anderen Controllern. Dafür sind die oberen drei Zustände (connect, ctrl_connect und disconnect) in Abbildung B.9 zuständig. Für die Weiterleitung der Signalisierungsnachrichten zwischen ATM und den Bearbeitungsprozessen für lokale und globale Nachrichten sind die unteren drei Zustände (from_global, from_local,

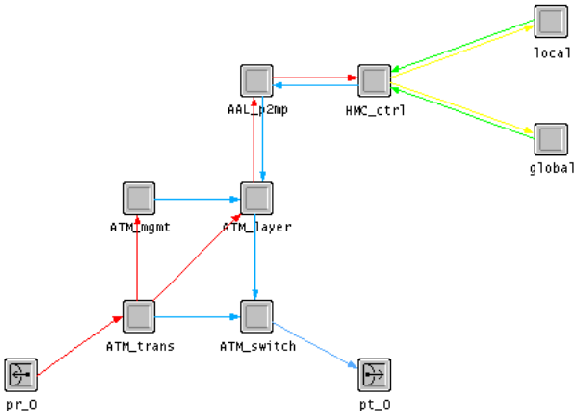


Abbildung B.8.: Knotenmodell des Controllers.

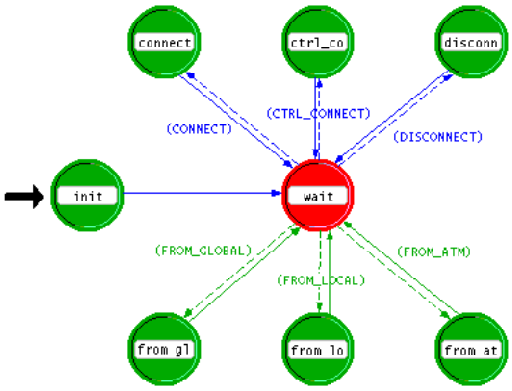


Abbildung B.9.: Prozessmodell des Verteilers im Controller.

from_atm) in Abbildung B.9 verantwortlich. Je nachdem auf welcher ATM-Verbindung die Nachrichten eingetroffen sind, können sie einem Prozess (lokal oder global) zugeordnet werden.

Die Behandlung der lokalen Gruppenkommunikation (Kapitel 5 ab Seite 65) ist durch das Prozessmodell aus Abbildung B.10 realisiert. Auf der rechten Seite in Abbildung B.10 sind vier Zustände (r_leave, s_leave, r_join und s_join), die für die Gruppenan- und -abmeldungen der lokalen Endsysteme verantwortlich sind. Der Zustand connect bearbeitet die Anmeldung lokaler MCS und der Zustand mcs_response wird verwendet, um Antworten der MCS auf Belastungsanfragen zu verarbeiten und daraufhin einen Sender einem MCS zuordnen zu können. Der Zustand MCS_change ist für die Verarbeitung von Nachrichten des Prozesses für die globale Kommunikation vorgesehen. Damit wird das Prinzip realisiert, dass entfernte MCS genauso wie lokale Gruppenteilnehmer behandelt werden.

Für die globale Gruppenkommunikation bei SkaGAN (Kapitel 6 ab Seite 85) ist der in Abbildung B.11 dargestellte Prozess zuständig. Den meisten Raum nimmt dabei die Bearbeitung von Nachrichten anderer Controller ein. Die Zustände update, update_low und update_high sind für die Bearbeitung der Updates zwischen den Controllern verant-

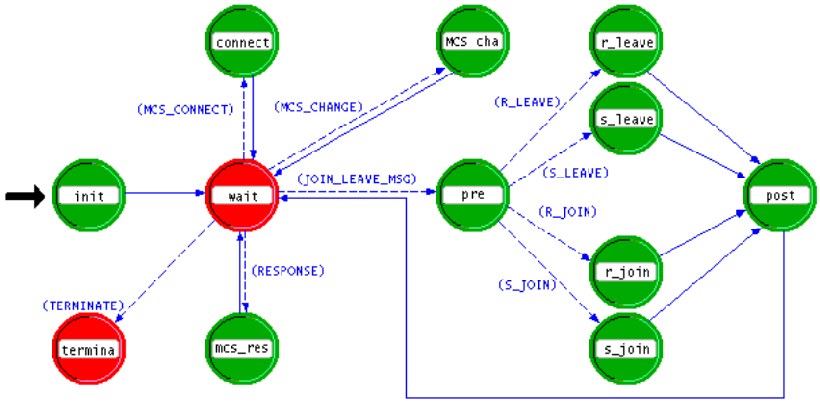


Abbildung B.10.: Prozessmodell der lokalen Gruppenkommunikation im Controller.

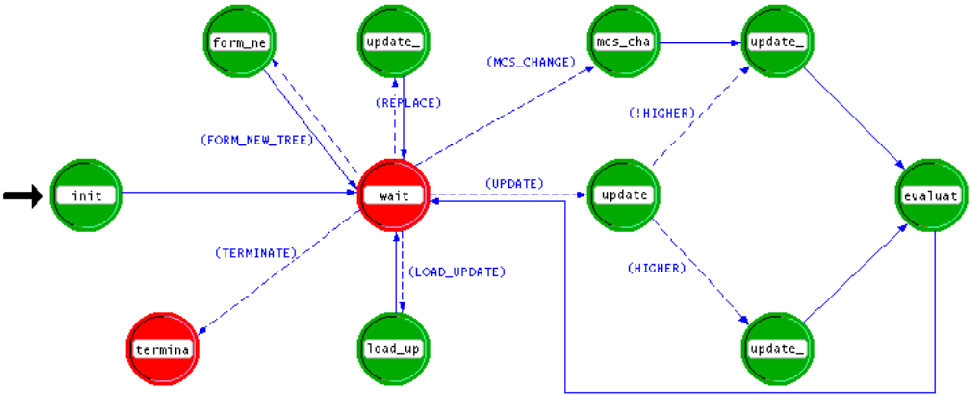


Abbildung B.11.: Prozessmodell der globalen Gruppenkommunikation im Controller.

wortlich. Eine anschließende Bereinigung des Datenbestandes wird im Zustand evaluate durchgeführt. Für die Berücksichtigung lokaler Gruppenänderungen ist der Zustand mcs_change vorhanden. Änderungen aus dem Prozess für die lokale Gruppenkommunikation (Abbildung B.10) werden hier mit in das globale Kommunikationsmodell integriert.

Die weiteren Zustände sind für die Erweiterungen (Unterkapitel 6.3 ab Seite 113) notwendig. Im Zustand load_update werden MCS-Belastungsmeldungen verarbeitet und evtl. eine MCS-Ersetzung oder der Aufbau eines neuen Baumes eingeleitet. Der Zustand update_replace ist für die Ersetzung eines MCS verantwortlich. Hier werden die notwendigen Daten und Nachrichten verwaltet. Für die andere Erweiterung, mehrere Bäume pro Gruppe, ist der Zustand form_new_tree vorhanden.

B.6. Grenzen der Simulation

Für die Simulation des SkaGAN-Ansatzes in einem Weitverkehrsnetz war es wichtig, möglichst große ATM-Netze heranzuziehen. Für jede ATM-Komponente erzeugt OpNet aber eine Instanz der ATM- und AAL-Schicht. Dies führt zu einem erheblichen Speicher-

verbrauch und begrenzt somit die simulierbare Netzwerkgröße. Der Simulationsrechner ist eine Ultra 60 von Sun Microsystems mit 1GByte RAM-Speicher. Es hat sich gezeigt, dass damit Simulationen mit maximal ca. 1000 ATM-Komponenten möglich sind.

C. Nachrichtenformate

Dieser Anhang enthält die Beschreibung aller verwendeten Signalisierungsnachrichten, deren Format und Verwendungszweck. Der grundsätzliche Aufbau aller Formate ist bereits in Unterkapitel 4.2.1, ab Seite 58 beschrieben worden. An das dort vorgestellte Schema schließt sich die hier folgende Beschreibung an. Zur besseren Übersichtlichkeit sind die Nachrichten in drei folgenden Bereiche untergliedert: Endsystem – Controller, MCS – Controller und Controller – Controller.

Die Nachrichten werden nach dem in Kapitel 4.2.1 beschriebenen Konzept dargestellt und die Bedeutung der einzelnen Felder werden erläutert.

C.1. Endsystem – Controller

Der Nachrichtenaustausch zwischen Endsystem und Controller dient hauptsächlich der An- und Abmeldung von Gruppenteilnehmern. Hierzu stehen vier Nachrichten zur Verfügung (Tabellen C.1, C.2, C.3 und C.4), mit denen sich ein Endsystem beim lokalen Controller für eine Gruppe als Sender oder Empfänger an- oder abmelden kann. Es ist nur die Angabe der Gruppenadresse und der Sichtbarkeit der Adresse notwendig, die Quelle, also das Endsystem wird vom Controller über die ATM-Verbindung ermittelt, auf der die Nachricht eingetroffen ist.

Darüber hinaus gibt es nur noch eine weitere Nachricht (Tabelle C.5), die einen Sender einem MCS zuweist. Diese Nachricht erhält ein neu zu einer Gruppe beigetretener Sender als Antwort auf eine **SJOIN**-Nachricht.

SJOIN		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, zu der der Sender beitreten möchte
Scope	Number	Sichtbarkeit der Gruppenadresse

Tabelle C.1.: Nachrichtenformat für die Senderanmeldung.

SLEAVE		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, aus der der Sender austreten möchte
Scope	Number	Sichtbarkeit der Gruppenadresse

Tabelle C.2.: Nachrichtenformat für die Senderabmeldung.

RJOIN		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, zu der der Empfänger beitreten möchte
Scope	Number	Sichtbarkeit der Gruppenadresse

Tabelle C.3.: Nachrichtenformat für die Empfängeranmeldung.

RLEAVE		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, aus der der Empfänger austreten möchte
Scope	Number	Sichtbarkeit der Gruppenadresse

Tabelle C.4.: Nachrichtenformat für die Empfängerabmeldung.

SET_MCS		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die der MCS bestimmt ist
MCS_Address	ATM_Address	Adresse des MCS, der für diesen Sender und diese Gruppe zuständig ist
MCS_SAP	SAP	Zur ATM-Adresse gehörige SAP-Kennung

Tabelle C.5.: Nachrichtenformat für die Zuordnung eines MCS zu einem Sender.

C.2. MCS – Controller

Die Nachrichten zwischen MCS und Controller betreffen zum einen die Zuordnung von Gruppenteilnehmern zu einem MCS und zum anderen die Meldung der MCS-Belastungen an den Controller.

Für das Hinzufügen oder das Entfernen eines Empfängers wird vom Controller die **MCS_UPDATE**-Nachricht (Tabelle C.6) an den MCS geschickt. Über das Flag Is_Upper wird zwischen einem lokalen Empfänger (oder untergeordneter MCS) und einem höheren MCS unterschieden. Das ist wichtig, da der höhere MCS über eine separate ATM-Verbindung angebunden wird. Die Sub_Group ist für den Einsatz mehrerer Bäume für eine Gruppe notwendig, damit der MCS zwischen den Bäumen unterscheiden kann.

Damit der Controller einen Sender einem MCS zuordnen kann, benötigt er die Belastungen aller MCS. Hier zu verschickt er die **REQUEST_LOAD**-Nachricht (Tabelle C.7) an alle MCS. Jeder MCS, der die Nachricht erhalten hat antwortet mit seinem zuletzt gemessenen Belastungswert (Tabelle C.8).

Für die Erweiterungen wird davon ausgegangen, dass die MCS aktiv ihre eigene Belastung überwachen und bei einer größeren Abweichung eine Meldung an den Controller

senden. Hierfür wird die Nachricht **LOAD_UPDATE** verwendet (Tabelle C.9), die wie die **RESPONSE_LOAD**-Nachricht nur ein Feld für den aktuellen Belastungswert zur Verfügung stellt.

MCS_UPDATE		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die der Empfänger bestimmt ist
Recv_Address	ATM_Address	Adresse des Empfängers, der diesem MCS zugeordnet worden ist
Recv_SAP	SAP	Zur ATM-Adresse gehörige SAP-Kennung
Add	Boolean	1 = zur Gruppe hinzufügen, 0 = entfernen
Is_Upper	Boolean	entfernter MCS ist 1 = höherer MCS, 0 = untergeordneter MCS
Sub_Group	Number	für parallele Bäume, enthält Baum-ID

Tabelle C.6.: Nachrichtenformat für die Zuordnung eines Empfängers zu einem MCS.

REQUEST_LOAD		
Name	IE	Beschreibung

Tabelle C.7.: Belastungsanfrage vom Controller an einen MCS.

RESPONSE_LOAD		
Name	IE	Beschreibung
Load	FixedFloat	MCS-Belastung im Bereich 0-1

Tabelle C.8.: Belastungsantwort vom MCS an den Controller.

LOAD_UPDATE		
Name	IE	Beschreibung
Load	FixedFloat	MCS-Belastung im Bereich 0-1

Tabelle C.9.: Belastungsmeldung vom MCS an den Controller.

C.3. Controller – Controller

Die Nachrichten zwischen Controllern sind für den Auf- und Abbau der Baumstrukturen notwendig. Für das Basisschema wird nur ein Nachrichtenformat benötigt, die **CTRL_UPDATE**-Nachricht (Tabelle C.10). Für eine Gruppe, die über Gruppenadresse, Sichtbarkeit und Untergruppe identifiziert wird, werden eine Reihe von Schaltern

übertragen, die den Zustand eines MCS im Baum beschreiben. Hierzu ist zunächst die Adresse des MCS angegeben, dessen Zustand sich geändert hat. Der Schalter Is_Local gibt an, ob der MCS dem sendenden Controller lokal zugeordnet ist oder nur über einen weiteren Controller erreicht werden kann. Der Schalter Is_Upper wird nur beachtet, wenn Is_Local = 0 ist, und gibt die relative Position des MCS im Vergleich zum sendenden Controller im Baum an. Die beiden Schalter Is_Send und Is_Recv geben an, ob der MCS Sender und/oder Empfänger in der Gruppe ist.

Die weiteren Nachrichten sind für die Erweiterungen, MCS-Ersetzung und mehrere Bäume pro Gruppe, notwendig. Für die MCS-Ersetzung ist der Auslöser die **MCS_REPLACE_REQUEST**-Nachricht. Damit meldet ein Controller an den höheren Controller, dass ein MCS aus der angegebenen Gruppe austreten soll. Die Nachricht enthält die Adresse des MCS, und ob der MCS Sender und/oder Empfänger in der Gruppe ist. Diese Information wird mitgeliefert, um damit einen neuen MCS besser auswählen zu können.

Für die Auswahl eines Ersatz-MCS wird vom zuständigen Controller die **MCS_LOAD_GLOBAL_REQUEST**-Nachricht (Tabelle C.12) an die untergeordneten Controller im Verwaltungsbaum versendet. Die Controller antworten mit der **MCS_LOAD_GLOBAL_RESPONSE**-Nachricht (Tabelle C.13), die für jeden in Frage kommenden MCS generiert wird und die MCS-Adresse und dessen Belastung enthält. Anhand dieser Antworten wird der Ersatz-MCS vom zuständigen Controller ausgewählt (hierfür kann wieder die **CTRL_UPDATE**-Nachricht verwendet werden).

Die Erweiterung für mehrere Bäume pro Gruppe benötigt vier weitere Nachrichten, die in den einzelnen Phasen bei der Etablierung eines neuen Baumes eingesetzt werden. Mit der **FORM_TREE_REQUEST**-Nachricht (Tabelle C.14) wird eine Anfrage von einem unteren Controller an einen höheren Controller gesendet, die den Aufbau eines neuen Baumes einleitet. Ausgelöst wird diese Anfrage durch eine zu hohe MCS-Belastung. Daraufhin versucht der höchste Controller in der Gruppe festzustellen, ob ausreichend MCS im ATM-Netz für einen weiteren Baum zur Verfügung stehen. Das geschieht mit den Nachrichten **MCS_QUERY_REQUEST** und **MCS_QUERY_RESPONSE** (Tabellen C.15 und C.16). Nur wenn in den zurücklaufenden **MCS_QUERY_RESPONSE**-Nachrichten ausreichend MCS gemeldet werden, wird der Aufbau eines neuen Baums mit der **FORM_NEW_TREE**-Nachricht (Tabelle C.17) eingeleitet. Diese Nachricht enthält keine Angaben über die zu verwenden Controller, diese Entscheidung wird den lokalen Controllern überlassen.

CTRL_UPDATE		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID
MCS_Address	ATM_Address	Adresse des MCS mit geänderten Zustand
Is_Local	Boolean	der MCS ist 1 = lokaler MCS, 0 = entfernter MCS
Is_Upper	Boolean	entfernter MCS ist 1 = höherer MCS, 0 = untergeordneter MCS
Is_Send	Boolean	1 = MCS ist Sender, 0 = MCS ist kein Sender
Is_Recv	Boolean	1 = MCS ist Empfänger, 0 = MCS ist kein Empfänger

Tabelle C.10.: Nachricht für Gruppenänderung zwischen Controllern.

MCS_REPLACE_REQUEST		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID
MCS_Address	ATM_Address	Adresse des MCS mit geänderten Zustand
Is_Send	Boolean	1 = MCS ist Sender, 0 = MCS ist kein Sender
Is_Recv	Boolean	1 = MCS ist Empfänger, 0 = MCS ist kein Empfänger

Tabelle C.11.: Belastungsmeldung vom MCS an den Controller.

MCS_LOAD_GLOBAL_REQUEST		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID

Tabelle C.12.: Belastungsanfrage vom Controller an entfernte MCS.

MCS_LOAD_GLOBAL_RESPONSE		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID
MCS_Address	ATM_Address	Adresse des Ersatz-MCS
Load	FixedFloat	Belastung des Ersatz-MCS im Bereich 0-1

Tabelle C.13.: Belastungsantwort vom entfernten MCS an den Controller.

FORM_TREE_REQUEST		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID

Tabelle C.14.: Anfrage nach Aufbau eines neuen Baums.

MCS_QUERY_REQUEST		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID

Tabelle C.15.: Anfrage der MCS-Anzahl vom höheren Controller.

MCS_QUERY_RESPONSE		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID
Nr_MCS	Integer	Anzahl der MCS im Unterbaum

Tabelle C.16.: Antwort mit der aktuellen MCS-Anzahl im Unterbaum.

FORM_NEW_TREE		
Name	IE	Beschreibung
Group_Address	IP_Address	Gruppe, für die die Nachricht bestimmt ist
Scope	Number	Sichtbarkeit der Gruppenadresse
Sub_Group	Number	für parallele Bäume, enthält Baum-ID

Tabelle C.17.: Anweisung einen neuen Baum aufzubauen.

D. Abkürzungsverzeichnis

AAL	ATM Adaption Layer
ABR	Available Bit Rate
ARP	Address Resolution Protocol
ATM	Asynchronous Transfer Mode
BOP	Beginning of Packet
BUS	Broadcast and Unknown Server
CBR	Constant Bit Rate
CBT	Core Based Tree
CCVC	Cluster Control VC
CLIP	Classical IP over ATM
CLP	Cell Loss Priority
CONGRESS	Connection-oriented Group Address Resolution Service
COP	Continuation of Packet
CRAM	Cell Relabeling at Merge-points
CRC	Cyclic Redundancy Check
CSCW	Computer Supported Cooperative Work
CTRL	Controller
DVMRP	Distance Vector Multicast Routing Protocol
EARTH	Easy IP Multicast Routing through ATM Clouds
EOP	End of Packet
FTP	File Transfer Protocol
ICMP	Internet Control Message Protocol
ID	Identifier
IDMR	Inter Domain Multicast Routing
IE	Information Element
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol
IMSS	IP Multicast Shortcut Service
IP	Internet Protocol
IP-SENATE	IP Multicast Service for Non-broadcast Access Networking Technology
ISDN	Integrated Services Digital Network
ITU-T	International Telecommunications Union - Telecommunications
LAN	Local Area Network
LANE	Local Area network Emulation
LIS	Logical IP Subnet

MARS	Multicast Address Resolution Server
MCS	Multicast Server
MID	Multiplexing Identifier
MLIS	Multicast Logical IP Subnet
MPOA	Multi-Protocol over ATM
NBMA	Non-Broadcast Multiple Access
NHRP	Next Hop Resolution Protocol
OSPF	Open Shortest Path First
PIM	Protocol Independant Multicast
PIM-SM	Protocol Independant Multicast - Sparse Mode
PNNI	Private Network to Network Interface
PVC	Permanent Virtual Circuit
QOS	Quality of Service
RM	Ressource Management
RPF	Reverse Path Forwarding
RSVP	Ressource Reservation Protocol
SAP	Service Access Point
SCSP	Server Cache Synchronisation Protocol
SCVC,	Server Control Virtual Connection
SEAM	Scalable and Efficient ATM Multicast
SG	Server Group
SID	Source Identifier
SMART	Shared Many-to-many ATM Reservations
SMS	Selective Multicast Server
SPAM	Simple Protocol for ATM Multicast
SSCOP	Service Specific Connection Oriented Protocol
SVC	Switched Virtual Circuit
SkaGAN	Skalierbare Gruppenkommunikation über ATM
TCP	Transmission Control Protocol
UBR	Unspecified Bit Rate
UNI	User Network Interface
VC	Virtual Channel
VCC	Virtual Channel Connection
VCI	Virtual Channel Identifier
VCL	Virtual Channel Link
VENUS	Very Extensive Non-Unicast Service
VP	Virtual Path
VPC	Virtual Path Connection
VPI	Virtual Path Identifier
WAN	Wide Area Network

E. Namenskonventionen

Bezeichnung der Datenpakete:

Paket	AAL5-Paket
Datenpaket	Paket einer Anwendung
Nachrichtenpaket/Nachricht	Paket der Signalisierung von SkaGAN

MCS Lagebezeichnungen:

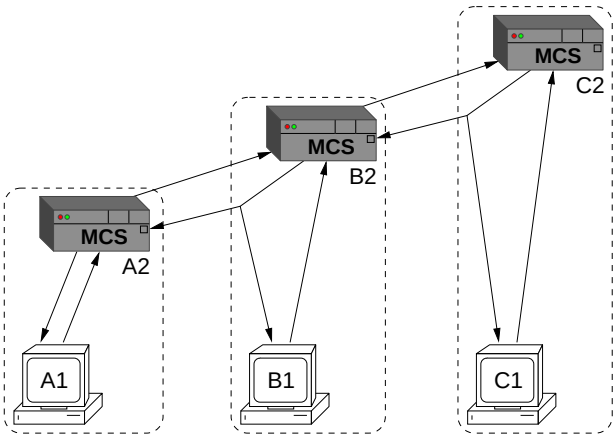


Abbildung E.1.: Positionen der MCS.

Anhand des Beispiels aus Abbildung E.1 sollen die verschiedenen Lagebezeichungen der MCS zueinander dargestellt werden:

- A2 ist lokaler (primärer) MCS bzgl. Endsystem A1.
- B2 ist primärer MCS von A2 und C2 ist primärer MCS von B2.
- B2 und C2 sind übergeordnete oder höhere MCS von A2.
- B2 ist der nächst höhere MCS von A2 und C2 ist der nächst höhere MCS von B2.
- A2 und B2 sind untergeordnete MCS von C2
- C2 ist entfernter (primärer) MCS von A2.

Literaturverzeichnis

[1] Wittmann, R. ; Zitterbart, M.: *Multicast – Protocols and Applications*. Morgan Kaufmann Publishers, 2000 (ISBN 1-55860-645-9)

[2] Miller, C. K.: *Multicast Networking and Applications*. Addison Wesley, 1998 (ISBN 0-201-30979-3)

[3] Deering, S.: Host Extensions for IP Multicasting / Internet Engineering Task Force. 1989 (1112). – Request for Comments

[4] Meyer, D.: Administratively Scoped IP Multicast / Internet Engineering Task Force. 1998 (2365). – Request for Comments

[5] Deering, Stephen E. ; Cheriton, David R.: Multicast routing in datagram internetworks and extended LANs. In: *Transactions on Computer Systems* 8 (1990), Mai, Nr. 2, S. 85–110

[6] ATM Forum. *ATM User-Network Interface (UNI) Signalling Specification Version 3.1*. ftp://ftp.atmforum.com/pub/approved-specs/af-uni-0010.002. September 1994

[7] Alles, A. *ATM Internetworking*. Internal Paper, Cisco Systems, Inc. Mai 1995

[8] Kyas, O.: *ATM-Netzwerke*. D - 50105 Bergheim : Datacom, 1995 (ISBN 3-89238-108-9)

[9] Smirnov, M.: EARTH - EAsy IP multicast Routing THrough ATM clouds / Internet Engineering Task Force. 1997. – Internet Draft, Expired

[10] Farinacci, D. ; Meyer, D. ; Rekhter, Y.: Intra-LIS IP multicast among routers over ATM using Sparse Mode PIM / Internet Engineering Task Force. 1998 (2337). – Request for Comments

[11] Anker, T. ; Breitgand, D. ; Dolev, D. ; Levy, Z.: IMSS: IP Multicast Shortcut Service / Internet Engineering Task Force. 1998. – Internet Draft

[12] Plummer, D.: An Ethernet Address Resolution Protocol - or - Converting Network Protocol Addresses to 48.bit Ethernet Address for Transmission on Ethernet Hardware / Internet Engineering Task Force. 1982 (826). – Request for Comments

[13] Estrin, D. ; Farinacci, D. ; Helmy, A. ; Thaler, D. ; Deering, S. ; Handley, M. ; Jacobson, V. ; Liu, C. ; Sharma, P. ; Wei, L.: Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification / Internet Engineering Task Force. 1998 (2362). – Request for Comments

[14] Deering, S. ; Estrin, D. ; Farinacci, D. ; Jacobson, V. ; Helmy, A. ; Meyer, D. ; Wei, L.: Protocol Independent Multicast Version 2 Dense Mode Specification / Internet Engineering Task Force. 1999. – Internet Draft. Work in progress

[15] Fenner, W.: Internet Group Management Protocol, Version 2 / Internet Engineering Task Force. 1997 (2236). – Request for Comments

[16] Postel, J.: Internet Control Message Protocol / Internet Engineering Task Force. 1981 (792). – Request for Comments

[17] Kumar, V.: *MBone: Interactive Multimedia On The Internet*. Macmillan Publishing (Simon & Schuster), 1995 (ISBN 1-56205-397-3)

[18] Black, Uyless D.: *ATM - Foundation for Broadband Networks*. Englewood Cliffs, New Jersey : Prentice Hall PTR, 1995 (ISBN 0-13-297178-X). – ISBN 0–13–297178–X

[19] *International Telecommunication Union – Telecommunication (ITU-T)*. http://www.itu.int/ITU-T

[20] ATM Forum. *Private Network-Network Interface Specification Version 1.0 (PNNI 1.0)*. ftp://ftp.atmforum.com/pub/approved-specs/af-pnni-0055.000.pdf. Juli 1996

[21] ATM Forum. *ATM User-Network Interface (UNI) Signalling Specification Version 4.0*. ftp://ftp.atmforum.com/pub/approved-specs/af-sig-0061.000.pdf. Juli 1996

[22] Black, Uyless D.: *ATM - Vol. II, Signalling in Broadband Networks*. Englewood Cliffs, New Jersey : Prentice Hall PTR, 1998 (ISBN 0-13-571837-6)

[23] Laubach, M. ; Halpern, J.: Classical IP and ARP over ATM / Internet Engineering Task Force. 1998 (2225). – Request for Comments

[24] ATM Forum. *LAN Emulation over ATM 1.0*. ftp://ftp.atmforum.com/pub/approved-specs/af-lane-0021-000.pdf. Januar 1995

[25] The ATM Forum. *Multi-Protocol Over ATM Version 1.0*. ftp://ftp.atmforum.com/pub/approved-specs/af-mpoa-0087.000.pdf. Juli 1997

[26] Heinanen, J.: Multiprotocol Encapsulation over ATM Adaption Layer 5 / Internet Engineering Task Force. 1993 (1483). – Request for Comments

[27] The ATM Forum. *Baseline Text for MPOA*. ATM Forum document atmf 95-0824. Juli 1995

[28] Armitage, G.: Support for Multicast over UNI 3.0/3.1 based ATM Networks / Internet Engineering Task Force. 1996 (2022). – Request for Comments

[29] Gauthier, E. ; Le Boudec, J. Y. ; Oechslin, P.: SMART: A Many-to-Many Multicast Protocol for ATM. In: *IEEE Journal on Selected Areas in Communications* 15 (1997), April, Nr. 3, S. 458–472

[30] Armitage, G.: Issues affecting MARS Cluster Size / Internet Engineering Task Force. 1997 (2121). – Request for Comments

[31] Grossglauser, M. ; Ramakrishnan, K. K.: SEAM: Scalable and Efficient ATM Multicast. In: *Proc. of IEEE INFOCOM*, 1997

[32] Armitage, G.: VENUS - Very Extensive Non-Unicast Service / Internet Engineering Task Force. 1997 (2191). – Request for Comments

[33] Komandur, S. ; Mossé, D.: SPAM: A Data Forwarding Model for Multipoint-to-Multipoint Connection Support in ATM Networks. In: *Proc. of 6th International Conference on Computer Communications and Networks*, IEEE Computer Society, September 1997

[34] Komandur, S. ; Crowcroft, J. ; Mossé, D.: CRAM: Cell Re-labeling at Merge-points for ATM Multicast. In: *Proc. of IEEE International Conference on ATM (ICATM)*, 1998

[35] Luciani, J. ; Gallo, A.: A Distributed MARS Service Using SCSP / Internet Engineering Task Force. 1998 (2443). – Request for Comments

[36] Talpade, R. ; Ammar, M.: Multicast Server Architectures for MARS-based ATM Multicasting / Internet Engineering Task Force. 1997 (2149). – Request for Comments

[37] Luciani, J. ; Katz, D. ; Piscitello, D. ; Cole, B. ; Doraswamy, N.: NBMA Next Hop Resolution Protocol (NHRP) / Internet Engineering Task Force. 1998 (2332). – Request for Comments

[38] Pusateri, T.: Distance Vector Multicast Routing Protocol / Internet Engineering Task Force. 2000. – Internet Draft

[39] Luciani, J. ; Armitage, G. ; Halpern, J. ; Doraswamy, N.: Server Cache Synchronization Protocol (SCSP) / Internet Engineering Task Force. 1998 (2334). – Request for Comments

[40] Moy, J.: OSPF Version 2 / Internet Engineering Task Force. 1998 (2328). – Request for Comments

[41] Talpade, R. ; Armitage, G. ; Ammar, M. H.: Experience with Architectures for Supporting IP Multicast over ATM. In: *Proceedings of the IEEE ATM'96 Workshop*, IEEE Computer Society, August 1996

[42] Talpade, R. ; Ammar, M. H.: Multicast Server Architectures for Supporting IP Multicast over ATM. In: *Proceedings of the 7th IFIP Conference on High Performance Networking*, 1997

[43] Zhong, W. D. ; Yukimatsu, K.: Design requirements and architectures for multicast ATM switching. In: *IEICE Trans. on Commun.* E77 (1994), November, Nr. B, S. 1420–1428

[44] Ballardie, A.: Core Based Trees (CBT version 2) Multicast Routing / Internet Engineering Task Force. 1997 (2189). – Request for Comments

[45] Ballardie, A.: Core Based Trees (CBT) Multicast Routing Architecture / Internet Engineering Task Force. 1997 (2201). – Request for Comments

[46] Dalal, Yogen K. ; Metcalfe, Robert M.: Reverse path forwarding of broadcast packets. In: *Communications of the ACM* 21 (1978), Dezember, Nr. 12, S. 1040–1048

[47] Franke, A.: *Abbildung von IP Integrated Services Dienstgüteparameter auf ATM Netzwerke*, Technische Universität Braunschweig, Studienarbeit, Oktober 1998

[48] Bouazizi, I.: *Multiparty-Kommunikation in ATM-Netzen; Implementation der Signalisierung*, Technische Universität Braunschweig, Studienarbeit, September 1999

[49] Afifi, H. ; Bonjour, D. ; Elloumi, O.: TCP over Non Exsistent IP for ATM Networks. In: *Proc. of Joint European Networking Conference (JENC)*, 1996

[50] Anwander, M.: *Entwurf und Simulation eines Konzeptes zur lokalen Gruppenkommunikation über ATM-Netzen*, Technische Universität Braunschweig, Diplomarbeit, August 2000

[51] Schubert, K.: *Heterogenes Multipeer in ATM-Netzen*, Technische Universität Braunschweig, Studienarbeit, Dezember 1998

[52] Böger, A. ; Zitterbart, M.: Towards Support for Heterogeneous Multipeer across ATM Networks. In: *Proceedings of 9th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN)*, 1998

[53] Stallings, W.: *High-Speed Networks.* Prentice Hall, 1998 (ISBN 0-13-525965-7)

[54] Almeroth, K. ; Ammar, M.: Collecting and Modeling the Join/Leave Behavior of Multicast Group Members in the MBone. In: *Proceedings of HDPC 96*, 1996, S. 209–216

[55] Grimsehl, P.: *Analyse und Modellierung von Gruppenkommunikationsverbindungen*, Technische Universität Braunschweig, Studienarbeit, Februar 2000

[56] Zipf, G.: *Human Behaviour and the Principle of Least Effort.* Addison Wesley, 1949

[57] Böger, A. ; Zitterbart, M.: Heterogeneous Group Communication over ATM Networks. In: *Proceedings of Networked Group Communication Workshop NGC'99, Poster Session*, 1999

[58] Böger, A. ; Zitterbart, M.: Signalling Support for Scalable Group Communication over ATM Networks. In: *Kommunikation in Verteilten Systemen (KiVS)*, 2001, S. 147–158

[59] Handley, M. ; Perkins, C. ; Whelan, E.: Session Announcement Protocol / Internet Engineering Task Force. 2000 (2974). – Request for Comments

[60] Fickel, T. B.: *Skalierbare Gruppenkommunikation in großen ATM-Netzen*, Technische Universität Braunschweig, Diplomarbeit, Februar 2001

[61] *Mbone-DE*. http://www.mbone.de

[62] Diot, Christophe ; Dabbous, Walid ; Crowcroft, Jon: Multipoint Communication: A Survey of Protocols, Functions, and Mechanisms. In: *IEEE Journal On Selected Areas In Communications* 15 (1997), April, Nr. 3, S. 277–290

[63] Hwang, F. K. ; Richards, D. S.: Steiner tree problems. In: *IEEE Networks* 22 (1992), Januar, S. 55–89

[64] Talpade, R. ; Ammar, M. H.: VC consumption in the MARS architecture for IP Multicast over ATM. In: *Position paper at Workshop on Integration of IP+ATM networks*, 1996

[65] MIL3 ; MIL3 (Hrsg.). *OPNET Manual*. http://www.mil3.com